

Measurement and Evaluation

in EDUCATION,
PSYCHOLOGY,
and GUIDANCE

Georgia Sachs Adams

Copyright © 1964

by Holt, Rinehart and Winston, Inc.

20.7.93

Printed in the United States of America

7095

3-11-26

ADA

July, 1966

Copyright © 1964 by Holt, Rinehart and Winston, Inc.

All Rights Reserved

Library of Congress Catalog Card Number: 64-21407

20089-0214

Printed in the United States of America

Preface

This volume is a new general textbook designed to serve students and educators who wish a more thorough study of the concepts involved in various aspects of measurement than is presented in the two earlier books by the author in collaboration with Dr. Torgerson. These books, *Measurement and Evaluation for the Elementary School Teacher* and *Measurement and Evaluation for the Secondary School Teacher*, emphasized a functional approach to evaluation and included chapters on measurement, diagnosis, and corrective instruction in each of the major fields. The present text does not replace either of these books.

Many texts on measurement devote only a few chapters to concepts, while the remaining sections are concerned with the applications of measurement in the schools. In this volume, almost every chapter focuses on the development of concepts. By attaining a thorough understanding of these concepts, students will be helped to select or develop measurement procedures appropriate to their purposes and to interpret measurement data with due respect for (1) the inevitable errors involved in any sampling procedure and (2) the limitations of indirect approaches typically used as efficient substitutes for direct study of behavior.

Although applications have not been neglected, they are usually presented for the purpose of helping students increase their understanding of concepts. That is, illustrative examples are given to help students learn to apply measurement techniques with discrimination in (1) improving the quality of their decision-making about students, and in (2) reaching more valid judgments about students' abilities, interests, and personality traits.

The following are examples of the ways in which this textbook has aimed at a high level of understanding of measurement concepts.

In Chapter 4, the many aspects of validity in measurement are presented in all their complexity. Students are given many examples to help clarify the four major types of validity and their significance for the decision-making processes in education.

In Chapters 5 and 6, students are shown that available tests cannot be neatly classified into aptitude and achievement tests; illustrations are given of tests that represent different degrees of saturation with the verbal-educational factor. Instead of presenting separate chapters on different kinds of aptitude tests, all types have been considered together in Chapter 6 so that the student can see their similarities and differences.

In Chapters 7 through 9 on interests and personality, a few published inventories are described, but they are presented as illustrative of different approaches to the complex problems of measurement in these areas.

In Chapters 10 and 11 on teacher-made tests, students are asked not only to examine the traditional types of test items but to explore the full range of objectives and test items that illustrate a taxonomy of the cognitive domain. Such an exploration may help teachers to measure progress toward objectives which have heretofore been neglected in their evaluation programs.

Chapter 12, on the measurement of skills outcomes, has not been limited to a survey of published tests. The author has attempted to develop an understanding of the concepts involved in appraising student performance in the skills of communication, homemaking, industrial arts, and other fields. The techniques for evaluating products, as well as those for appraising "performance in process" are considered. Measurement of skills outcomes is neglected in the typical textbook.

Statistics are introduced throughout the textbook as needed in meaningful problem situations. The emphasis is on concept development rather than on the development of computational skills. However, the professor who wishes to introduce more instruction in statistics into the measurement course will find additional materials included in the Appendix.

Since the author believes that students taking a course in measurement should examine and appraise tests in their own subject fields, she has included in the Appendix a comprehensive, classified list of available tests, with references to critical reviews in the *Buros yearbooks*. This list should be of considerable value in project assignments.

The typical measurement class is heterogeneous with respect to the students' readiness for the study of measurement, as well as with respect to their needs within this instructional area. The typical class includes undergraduate students and experienced teachers and administrators, as well as students preparing for work as counselors, psychometrists, or school psychologists. For this reason, the author has attempted to include materials for those interested in individual testing, vocational interest and aptitude testing, various approaches to personality assessment, as well as group testing and educational diagnosis. Moreover, she has attempted to keep the textbook readable for the average student, while meeting the needs of advanced students through special tables, footnotes, and selected references.

The author wishes to express her appreciation to Dr. T. L. Torgerson,

who served as consultant throughout the writing and revision of the manuscript; to two coworkers at California State College at Los Angeles, Dr. Edwin Wandt and Dr. Carleton Shay, for their critical review of several chapters; to Dr. Miriam Bryan of Educational Testing Service for her review of Chapters 11 and 13. The author would also like to acknowledge her indebtedness to Mrs. Johanna de Graff for her highly competent assistance and her conscientious attention to detail in the typing of the manuscript. Finally, the patience and cooperation of her husband and children are gratefully acknowledged.

Georgia Sachs Adams
Professor of Education

Los Angeles
May 1964

Contents

PREFACE V

PART ONE

Basic Principles and Procedures

1. *Introduction* 3
Comparison of the Terms "Measurement" and "Evaluation" • The Basic Requirements of the Evaluative Process • The Teacher's Role in Evaluation • Illustrative Problems in Measurement and Evaluation • Summary Statement • Selected References • Discussion Questions and Suggested Activities

2. *Interpreting Test Data in Terms of Converted Scores* 17
Converted Scores Based on Comparison with a Perfect Score • Converted Scores Based on Comparisons Among Examinees • Defining Norming Populations and Selecting Norm Samples • Use of Different Types of Converted Scores in Computation • Summary Statement • Selected References • Discussion Questions and Suggested Activities

3. *Reliability* 68
Interpreting Test Scores in Terms of Sources of Variance • Computing Correlation Coefficients • Comparison of Standard Errors and Reliability Coefficients as Measures of Reliability • Methods of Estimating Reliability of Test Scores • Reliability of Difference Scores • Factors Affecting the Size of Reliability Coefficients • Improving the Reliability of Test Scores • Summary Statement • Selected References • Discussion Questions and Suggested Activities

4. *Validity* 103
Tests as Direct or Indirect Measures of Criterion Behavior •
Types of Judgments Made on the Basis of Test Results • Content
Validity • Concurrent Validity • Predictive Validity • Construct
Validity • Summary Statement • Selected References • Discus-
sion Questions and Suggested Activities
5. *Application of the Principles of Measurement
In the Selection of Tests* 149
Types of Tests Available • Evaluating Tests for Use for Specific
Purposes • Illustrative Use of the Summary Form with a
Standardized Test • Sources of Information about Published
Tests • Summary Statement • Selected References • Discussion
Questions and Suggested Activities

PART TWO

The Study of Individuals

6. *The Measurement of Aptitudes* 181
The Concepts of Aptitude and Achievement • Tests of General
Mental Ability or Scholastic Aptitude • Multiscore Tests of
Mental Abilities and Aptitude Test Batteries • Tests of Special
Aptitudes • Prognostic Tests • Purposes for Which Aptitude
Tests Are Used • Interpretation of Results from Scholastic
Aptitude Tests • Summary Statement • Selected References •
Discussion Questions and Suggested Activities
7. *The Measurement of Interests and Attitudes* 228
The Nature of Interests • Types of Interest Inventories • Basic
Interest Groups • Validity of Interest Inventories • Interpretation
of Interest-Inventory Results • Measurement of Attitudes •
Summary Statement • Selected References • Discussion Ques-
tions and Suggested Activities

8. *Informal Methods of Studying Personal-Social Adjustment* 257

The Nature of Personal-Social Adjustment • Personality Description • Sources of Data About the Personal-Social Adjustment of Individuals • Self-Report Techniques • Observation of Behavior • Obtaining the Opinions of Others: Teacher Rating Scales • Obtaining the Opinions of Others: Sociometric Techniques • Summary Statement • Selected References • Discussion Questions and Suggested Activities

9. *Personality Inventories and Projective Techniques* 295

The Psychometric Approach: Personality Inventories • Projective Techniques • Summary Statement • Selected References • Discussion Questions and Suggested Activities

PART THREE

The Improvement of Instruction

10. *Development, Try-out, and Revision of Teacher-made tests* 321

Importance of Teacher-made Tests • Characteristics of a Good Teacher-made Test • Planning Tests for Greater Content Validity • The Advantages and Disadvantages of Essay and Objective Tests • The Construction of Test Items • Evaluating Objective Teacher-made Tests • Preparing a Teacher-made Test for Use • Statistical Analysis of Test Results • Teacher Cooperation in Test Development • Providing Leadership in the Development of Teacher-made Tests and Other Aids to Evaluation • Summary Statement • Selected References • Discussion Questions and Suggested Activities

11. *The Taxonomy of Educational Objectives and Test Items Illustrative of Its Major Categories* 363

1.00 Knowledge • 2.00 Comprehension • 3.00 Application • 4.00 Analysis • 5.00 Synthesis • 6.00 Evaluation • Summary Statement • Selected References • Discussion Questions and Suggested Activities

12. *Evaluating Student Performance in the Skills* 401
Developing Tests of Skills Outcomes • Scoring Processes and Products • Illustrative Evaluation Techniques in the Communication, Manipulative, and Athletic Skills • Validity and Reliability of Evaluations of Student Performance in the Skills • Summary Statement • Selected References • Discussion Questions and Suggested Activities
13. *The Place of Standardized Achievement Tests in the Improvement of Instruction* 428
History of Achievement Testing • Uses of Standardized Achievement Tests • Leading Achievement Tests • Interpretation of Data from Achievement Testing Programs • Large-Scale Testing Programs • Summary Statement • Selected References • Discussion Questions and Suggested Activities
14. *Educational Diagnosis* 458
Measurement as Basic to Educational Diagnosis and Individualized Instruction • Levels of Diagnosis • Steps in Educational Diagnosis • Group Diagnosis • Basic Principles of Corrective Instruction • Summary Statement • Selected References • Discussion Questions and Suggested Activities

PART FOUR

Administrative, Supervisory, and Guidance Aspects of Measurement and Evaluation

15. *Planning and Administering the Evaluation Program* 487
Functions of the Evaluation Program • Characteristics of an Effective Evaluation Program • Planning the Evaluation Program • Guidance Workers and Psychologists as Resource Persons • Planning the Testing Program • Administering the Testing Program • Summary Statement • Selected References • Discussion Questions and Suggested Activities

16. *Summarizing, Recording, and Reporting Data about Individual Students* 507

Summarizing and Recording Data • Reporting Data to Students and Parents • Improving the Validity, Reliability and Comparability of Teachers' Marks • Summary Statement • Selected References • Discussion Questions and Suggested Activities

17. *Using Measurement Data in Individual and Group Guidance* 534

Guidance Responsibilities of Counselors and Teachers • Issues and Principles Involved in the Use of Measurement Data in Guidance • Guidance in Educational and Vocational Planning • Combining Group and Individual Approaches in Helping High School Students in Self-Appraisal and Life Planning • Summary Statement • Selected References • Discussion Questions and Suggested Activities

APPENDIXES 565

INDEX 641

PART ONE

The
Evaluative
Process

For many students, this textbook will constitute their first introduction to measurement and evaluation. For this reason, we have tried to develop the basic concepts in measurement through the use of nonmathematical explanations and realistic examples. Other students using this textbook will have already had a unit in measurement as part of an introductory course in education or psychology. These students will find that this textbook reviews concepts they have studied but leads them on to a higher level of understanding. In fact, through the study of summary tables, footnotes, and chapter references, they will be able to pursue their interest in any topic beyond the limits of the textbook. They will find that the concepts of measurement and evaluation have implications for almost every aspect of teaching, guidance, and administrative work.

EVERYDAY USES OF MEASUREMENT AND EVALUATION

During the course of a school day, teachers, principals, and other school personnel make many decisions about students and help them to make many decisions for themselves.

Decisions are best made on the basis of a good deal of information, and schools have cumulated considerable information about each student. The data that are helpful in grouping students for physical education, however, are not identical with the data that would be most helpful in deciding which students should be placed in accelerated classes in mathematics. We want data that are most relevant to each decision.

Decisions usually involve prediction. The following questions are typical: On which athletic team or in which mathematics class is this student likely to make the greatest growth? Is this student likely to be admitted to the col-

lege he plans to attend? The first of these two questions involves institutional decisions about students; the second involves providing data to the individual that will help him in making his own decision. In each case, we need relevant data, that is, data that will increase the accuracy of the judgments and inferences we make.

In order to improve the decision-making process, we need measurement data about individuals. For decisions regarding grouping in physical education, for example, we need to measure height, weight, and the student's achievement in certain physical skills. When we measure height or weight, our only sources of error would be inaccuracy in the instrument of measurement and careless errors in reading the results. When we measure achievement in the skills, however, we face new problems. The person's speed of running varies somewhat from time to time; when we time him on one occasion, we obtain only a sampling of his running ability. We must recognize that there are fluctuations in individual performance. When we make an inference from one sample of a boy's running ability, sampling error is involved. The only way in which we could determine the amount of variation in running time from one sampling to another would be to check the variations in performance. The variation might be less with older students than younger ones; it might be less with trained runners than with typical students.

When we measure such a skill as ability to throw baskets, we realize that we must standardize the testing conditions regarding distance from and height of the basket. When we measure batting skill, we are likely to find still greater variation in student performance from sample to sample because of the introduction of a new source of variation—the performance of the pitcher. Moreover, there is more subjectivity involved in scoring batting performance than there was in scoring “baskets.” In making such subjective judgments, coaches agree with each other much more than would untrained observers.

If we try to judge how well a student knows his spelling, arithmetic, or geography, we will get variations in his scores from one test to another because each test includes a different sampling of questions. From these examples, we see that (1) it is desirable to compare the performance of individuals under standard conditions and (2) we need to know the sources of error and the amount of error involved when we make judgments about an individual on the basis of a sampling of his behavior.

Obtaining summary scores that can be recorded, combined, and interpreted is simple for some dimensions of the individual and extremely difficult for others. We have no problem with height and weight; number of baskets thrown can be objectively scored; umpires are given training in judging objectively “hits,” “fouls,” “strikes,” and “balls.” It is no problem

to obtain "number right" or "percentage correct" on a test of spelling, history, or geography.

For decisions on selection of students for an accelerated class or the classification of students into various groups, number right or rank within the total group may suffice. For other uses of measurement, we would like to know whether students have achieved as well, or made as much progress in a year, as other students of their age and grade. To answer such questions, we need data concerning representative samples of students; we need to obtain age or grade norms or other types of data that aid in making meaningful comparisons of individuals and groups, and meaningful intra-individual comparisons, such as inferring that a student performs more adequately in mathematics than in social studies. Actually, when a teacher makes statements about a student behaving immaturely or acting like a younger child, he is interpreting a sample of the child's present behavior in terms of his own "norms," that is, his cumulated observational data about students in various age groups. The sampling of students he has observed, however, may or may not have been representative.

COMPARISON OF THE TERMS "MEASUREMENT" AND "EVALUATION"

When a representative of the American Psychological Association was asked to define "psychological tests" at a Congressional hearing, he gave a definition that, although broad, includes all essential characteristics. His definition was: "Psychological tests are nothing more than careful observations of actual performance under standard conditions."¹ The term "careful" implies that the procedures for sampling the performance and obtaining a record of it are systematic and objective enough that different observers would obtain reasonably comparable findings.

This definition could also be used for the concept of measurement. When we obtain measures (other than such physical measures as height and weight), we obtain and record data on a sampling of performance under standard conditions. "Evaluation" goes beyond measurement in that value judgments are involved. We measure a student's abilities in different areas. When we interpret these scores in terms of standards for his grade, in terms of his educational or vocational plans, or some other basis for making value judgments, we are no longer restricting ourselves to "measurement"; we are now "evaluating" his abilities or his progress. We

¹ "Report of Testimony at a Congressional Hearing," *American Psychologist*, vol. 13 (May 1958), 217-223.

measure a student's height; we evaluate it in terms of his goals (as jockey or high jumper). We measure a student's speed of reading; we evaluate it as unsatisfactory or satisfactory in terms of his age, previous experience, and educational goals.

Sometimes measurement and evaluation seem to be inseparable. For example, it is difficult to compare two poems, two short stories, two samples of handwriting, or two swimming strokes without rating them in terms of value judgments. We could measure poems with respect to number of words or length of line; our comparisons would be highly objective and consistent, but they would not be relevant to the goals of education. Hence, we choose to make a fairly subjective evaluation of the poems, rather than to measure them with respect to irrelevant dimensions.

THE BASIC REQUIREMENTS OF THE EVALUATIVE PROCESS

Since educators and psychologists are concerned chiefly with those measurements that can provide a basis for evaluation, it is well to get an overview of the steps in any evaluative process. Although most of our examples are taken from tests, in the narrower sense of the term, the same processes apply to observation of behavior as a basis for ratings on selected characteristics.

1. Determining what we wish to evaluate. The information we obtain should be *relevant* to the type of judgment we wish to make (for example, determining the level of children's reading vocabulary as a basis for selecting textbooks, or their speed of reading as a basis for judging length of reading assignments).
2. Defining what we wish to evaluate in terms of behavior. We need to define level of reading vocabulary in behavioral terms (for example, ability to choose the correct synonyms for words used in context). We also need to specify the group of words about which we wish to make inferences; if this is a fourth-grade class, we may wish to judge student knowledge of words included in third-, fourth-, and fifth-grade readers.
3. Selecting appropriate situations in which to observe performance. Here we would be concerned with selecting classroom or playground situations in which to observe behaviors, or types of test items that would elicit the behavior in which we are interested. Observation samples, or samplings of test items, should be sufficiently large and representative so that inferences based on the student's performance on the sample would give a fairly accurate indication of his usual level of performance, for example, his vocabulary knowledge on all words at specified grade levels, or of his characteristic reading speed. In a speed-of-reading test, we would like to sample the different kinds of textbook-type material that children are expected to read at that grade level, rather than having the test composed entirely of science materials, which some students would read at a rate above, and others at a rate below, their characteristic reading rate.

4. Getting a record. In paper-and-pencil tests the student provides his own record, which can be scored later. In the testing of physical skills, one might decide to record the performance on film; in the testing of pronunciation in foreign language, a tape recording might be utilized. In the evaluation of student characteristics (defined in behavioral terms), the teacher's observations could be recorded in narrative form, with emphasis on the description of behavior and the avoidance of judgmental terms.
5. Summarizing the evidence. As already explained, some evidence is easily summarized, for example, running time (in seconds), number of baskets made, and the like. In a vocabulary test, the score could be number of words for which the correct definition is selected; or we might decide to penalize the student for "incorrect guesses." Greater care is required in deciding how to summarize the data from a film on physical performance, a tape recording of a student's oral language performance, or a narrative report on children's behavior. We need to decide on the aspects to be "scored," the units to be used in scoring, and the like.

The examples of measurement we have given in this chapter have been fairly simple ones. How much more complex it is to attempt to measure a child's mental ability or aptitude for school work, his readiness for reading, his comprehension of the principle of photosynthesis, his ability to interpret maps and graphs, his ability to communicate effectively through written or oral expression, his interest in different vocations, or his attitude toward school.

Dyer feels that we must face up to the inherent complexity of educational measurement and work on its problems with humility, persistence, and all the competence we can muster.

I don't think the business of educational measurement is inherently simple . . . Any way you look at it, the measurement of human behavior is bound to be a terribly complex process, since the phenomena of human behavior are themselves as complex as anything in the universe. . . .

. . . teachers [must] more clearly realize the fact that measurement in one form or another is not only an indispensable part of their job but that, like all the other parts, it is full of difficulties and unanswered questions which require constant study and hard thinking and a willingness to move ahead on the basis of highly tentative hypotheses. . . .

On the one hand, we want them to become keenly aware of all the uncertainties in even the most careful measurement of pupil performance, and on the other hand, we want them to regard measurement as part and parcel of effective instruction. . . .

I think we can do it by getting them to regard all their classroom work as a continuous series of both minute and longer range experiments in pupil learning . . . Always central is the nagging question: "Did it work?" And this, of course, is where measurement gets into the act. . . . Of course, she can never know for certain whether her strategy is on the right track, since the instru-

ments and techniques on which she has to depend for checking her hunches and hypotheses about procedures are never wholly reliable or relevant. In most cases, I believe, the realization of uncertainty is achieved only after the cold steel of such ideas as sampling error, the variability of human behavior, and the fallibility of casual observation and personal judgment has entered the teacher's soul. She will learn to be sure that she cannot be sure, and accordingly her approach to the instructional task will become less rigid, more tentative and . . . more responsive to the individual learning needs of the pupils with whom she is confronted.²

THE TEACHER'S ROLE IN EVALUATION

Under the survey-testing approach to evaluation, which dominated the 1920s, the teacher had little say in the selection of the tests used or the interpretation of test results. Subject-matter content and standards of achievement were centrally established, and testing was considered an administrative-supervisory function.

The child-study approach, individualization of instruction, and the greater breadth and more functional nature of the modern curriculum have all modified the concept of evaluation. That is, the instructional and child-study uses of measurement and evaluation have tended to predominate in importance over the administrative-supervisory uses. Hence, the teacher now has a key role in the evaluation process.

As the teacher has achieved a more significant role in planning the educational experiences for his class, he has become responsible for appraising the worthwhileness of those experiences—the extent to which students are achieving educational goals. Furthermore, as the schools have committed themselves to the ideal of individualizing instruction and meeting student needs, so that all may learn more effectively, emphasis has been placed on accumulating and interpreting measurement data for *individuals* rather than for groups. Here again, because of his daily opportunities to obtain additional information about his students and to put the measurement data to use, the teacher is the key person.

The teacher's job as an evaluator has two essential facets, neither of which can occupy his exclusive attention. One has its orientation in the group instructional program and the extent to which the class, as a whole, is achieving its goals; the other has its orientation in the study of individual

² Henry S. Dyer, "What Point of View Should Teachers Have Concerning the Role of Measurement in Education," *15th Yearbook, National Council on Measurements Used in Education* (New York: The Council, 1958), pp. 11–12.

students—in diagnosis of their growth lags and discovery of the important causal factors for such lags. Both approaches are significant aspects of evaluation and indispensable to good teaching.

Although all of this textbook has been written with the objective of helping teachers to become more effective in their work in measurement and evaluation, Part Three is especially concerned with the contributions that can be made to the improvement of instruction.

ILLUSTRATIVE PROBLEMS IN MEASUREMENT AND EVALUATION

Measurement is an essential but difficult aspect of the educational program. If teachers are to persevere in becoming more informed and effective in this aspect of their teaching role, they need leadership, encouragement, and resource materials from their school administrator. Moreover, the school administrator has responsibilities to students, parents, and the community, which he cannot discharge effectively without the use of measurement techniques.

Instead of listing administrative responsibilities in evaluation, let us examine some of the problems faced by a hypothetical junior high school principal and his staff as they begin a new school year. These problems will be used as a basis for later discussion of the development of local norms, the basic concepts of reliability and validity, and many other concepts.

Let us assume that Mr. Smith has met with his staff to list problems they will face during the school year. Mr. Smith sensed that he and his staff members were making many important decisions without having adequate information. He felt sure that test results could be used in many cases to increase the informational basis for their decision-making. Many published tests were administered in his school; but the results, he suspected, were often filed away and forgotten. Some teachers felt that tests were of little value; others had an unquestioning reverence for test scores, which did not lead to desirable caution in their interpretation.

1. The first problem Mr. Smith raised was very important, yet very complex. He emphasized that, as principal, it was *his* obligation to find out, as best he could, whether students were making progress toward the major objectives of the educational program. He recognized the rights of individual teachers to decide on the specific learning activities and content through which important principles and concepts should be taught. He felt that it was clearly the teachers' responsibility to measure in their own way how accurately students had learned

these specifics. How could he best discharge his basic responsibility without interfering with the freedom of teachers as professional workers to plan the day-by-day learning experiences of their students? Staff discussion showed that answering such a question involved (a) careful selection of tests that would measure growth toward the major goals of education, rather than memory of details and (b) wisdom in the way in which test results were interpreted and used.

One staff member cautioned the principal about interpreting the comparative results for different teachers, not only because of differences between classes in average scholastic aptitude and background skills but also because of the difficulties inherent in measuring progress toward the less tangible objectives. Mr. Smith reassured his staff members that no one was asking him to *rank* his teachers in the order of their competence. His goal would be to develop hypotheses concerning classes that *appeared* to be doing a less-than-adequate job and then to test out these hypotheses through observation of teaching and a study of instructional and evaluation materials used. The assistance of supervisors and department heads would be needed, not only in checking upon his hypotheses concerning possible areas of weakness but in taking steps to help improve student achievement in any such areas.

2. Another problem was concerned with the wide variation in grading practices. Teachers make daily decisions concerning grades on homework, quizzes, products made in class, student participation in committee activities, and other aspects of student work. These data are summarized in grades that go home to parents, are recorded on students' cumulative records, and are used in making many decisions about students. When grades serve as the basis for decisions regarding eligibility and scholarships, when they are used by the student and his counselor as a partial basis for judging strengths and needs, and when they are used by schools and colleges as an aid in admissions and grouping, the assumption is made that grades are reasonably comparable.

The staff discussed what they could do to help teachers improve the comparability of their grades, at least *within* subject fields. The counselor gave the example of John and Peter, identical twins, both college-bound, whose grades in the past had approximated a B+ average. Certainly it seemed indefensible when a "trick of fate" resulted in John's being assigned to a history class in which he received an "easy A" because the teacher's standards of grading were generous and the competition from classmates was minimal, while Peter received a C in a class where competition was keen and the teacher was firmly committed to a policy of no A's and few B's.

Mr. Smith hesitated to interfere; yet, as principal, he knew that he and his staff should help teachers to do more fairly and effectively this job of evaluation and grading that was so important to students. Considerable opposition was expressed to requiring teachers to "grade on the curve" in every class. Agreement was reached, however, on two principles:

- a. If students' grades were to be assigned fairly, they must be assigned on the basis of students' ranks with respect to some composite score that reflected their many achievements.
- b. These composite scores should not be arrived at in an intuitive manner but through such procedures that the teacher could explain his basis for grading to students and parents.

3. A third problem was concerned with the advisability of using teaching machines to aid students in their learning of content and skills. One of his seventh-grade teachers was eager to try out their use in the teaching of spelling; the school district had agreed to supply the machines; the teacher, with the help of his coworkers, planned to compare results in the classes using teaching machines with those in matched groups, comparable in scholastic aptitude and other important factors, which were taught spelling in the usual manner. Fortunately, since the teacher was doing the research project for his master's degree, the planning of the research could be left largely to him and to his supervising professor at the college. However, Mr. Smith and his staff would be concerned with the findings of this study and the implications for their school, and hence they did have the responsibility of seeing that adequate tests of proficiency in spelling were selected to evaluate student progress. Certainly a test developed for nation-wide use would not be as adequate a basis for judgment as tests based on the state spellers.

4. Mr. Smith reported that he had been asked to head a city-wide evaluation committee. This group was to provide information to the superintendent and governing board that would help them to decide whether eighth-grade students were achieving an adequate knowledge of the history of their nation and their state. With respect to United States history, Mr. Smith realized that the committee's problem would be one of selecting a test, with national norms, that would be considered fair by students and teachers and that would have a sufficient range of difficulty to measure student knowledge in all the schools of a heterogeneous community.

The decision concerning the adequacy of student knowledge of state history posed a different problem. No published tests were available. It would probably be best for his committee to ask representative teachers from the different schools to plan a blueprint for the test, which would indicate how much emphasis should be given to different objectives and areas of content. The course of study and the state textbook could be analyzed as an aid to teacher judgments. Once decisions regarding objectives and content were made, however, one or more teachers with special proficiency in item writing could devise items that would be reviewed by the representative group.

Mr. Smith wondered if, as an additional by-product, he could obtain information that would help on one aspect of Problem 1, which would help him in appraising how well his staff was teaching American and state history. One

teacher reminded him that a test which might be quite adequate in sampling student *information* about history might fail to measure other important outcomes. In other words, a test that might serve to assure the public that a minimum competency in knowledge was being achieved by students would need to be supplemented by other measures before an over-all appraisal could be made concerning the adequacy of the instructional program.

5. Another problem concerned the selection of students for participation in a second-semester pilot program for gifted seventh-graders. These students would have their English, social studies, and mathematics instruction with a teaching team of three teachers especially interested in teaching the gifted and presumably most qualified to do so. Approximately one hundred students were to be selected from a seventh-grade population of five hundred. A committee had already decided that the selection should be made on the basis of the students' scholastic aptitude, reading ability, previous achievement in elementary school subjects, and teacher ratings on such significant traits as "seriousness of purpose."

Fortunately relevant data were already available in the school records. Moreover, the Committee on Gifted Students had already decided what factors to consider. An intelligence quotient of at least 120 and reading ability at the eighth-grade level or above were to be required. In choosing among the students who qualified on these two bases, considerable weight was to be assigned to sixth-grade marks. Greatest weight was to be given to ratings on "seriousness of purpose," "responsibility," and "originality," because the future of the entire program depended on how well students achieved in this pilot project.

Obtaining comparable data on such traits as the "seriousness of purpose" of students coming from 18 sixth-grade classes in 4 elementary schools seemed impossible. Staff members warned that the subject grades of teachers from the different elementary schools would not be comparable. Moreover, the Committee on Gifted Students had evidently assigned the greatest weight to the factors that would be the most difficult to measure, that is, teacher ratings on "originality" and other traits.

6. Decisions of still another type needed to be made in the counseling of eighth-grade students concerning their choice of college preparatory subjects for the ninth grade. This problem was similar to problem 5 in that *predictions* of future achievement were involved.

In problem 5 there was a maximum number that could be accommodated, hence this was a *selection* problem similar to that of a company selecting persons for a limited number of jobs. In problem 6, the principal is under no serious restrictions with respect to numbers; for example, if another class in algebra is needed, it can easily be substituted for one in general mathematics. The prediction problem here is one of classification or *placement* of students in terms of predicted "best treatment," as when a personnel manager recom-

mends the most suitable jobs for each of a group of persons who have already been employed.

Problems 5 and 6 differ in another respect. Problem 5 involves institutional decisions *about* students; problem 6 involves decisions *by* students. The job of the school is to help students and their parents make thoughtful choices, on the basis of interpretation of information most relevant to this decision. One of three choices could be made. For some students, the data on test scores and grades would show high academic aptitude and evidence of adequate motivation; for these students, taking *both* college-preparatory mathematics and language in ninth grade would be clearly indicated. The test scores and grades of other students might support the decision that it was best to attempt *neither* subject at the ninth-grade level, but to decide later (after a year of general mathematics) whether algebra was advisable. Still other students might find it advisable, on the basis of marginal or inconsistent data, to attempt *one but not both* college-preparatory subjects in the ninth grade.

7. The school counselors and the ninth-grade social studies teachers would be faced with the job of helping pupils make tentative choices regarding their vocational plans. Although *specific* decisions could be postponed until senior high school, the combination of the ninth-grade vocations unit and individual counseling should help students to appraise their abilities, achievements, and interests in such a way as to make wiser vocational choices. One of the counselors reported that on Career Day 90 percent of the ninth-grade students had signed up for a dozen popular vocations. He contended that the use and interpretation of an interest inventory would help many of them to broaden their perspectives. He also suggested that an aptitude test battery be administered, which would help him as counselor in conferences with students and parents on educational and vocational planning.

Problem 6 had required only a prediction of the *general level* of the student's success in college-preparatory subjects. In vocational guidance, however, the predictions are of a higher order of complexity. For example, would John, an able student who is trying to decide between two vocations that interest him, be *more* likely to succeed in engineering or medicine? Answering questions of this type requires comparative or *differential prediction*. Tests that help to predict a student's *general level* of achievement in academic work might not be adequate for predicting the *difference* between his predicted achievements in two vocational fields or two college curricula.

The student is concerned not only about his relative success in different vocational fields but also about how satisfying different vocations would be in the light of his interests and temperament. Making inferences concerning probable job satisfactions is unusually precarious. In predicting achievement, one is usually safe in assuming that the more ability a student has shown to date (in music, sports, or some other field), the more success he will probably

have in the future. That is, the higher the score the better the prospects. But in predicting job satisfaction, even this assumption seems questionable. For example, there might be an *optimum range* of intelligence for file clerks, with the brighter girls finding the field too routine and unchallenging; similarly, it might be best for a prospective librarian to be *below* average in certain characteristics, such as energy level and extroversion. Certainly the problem of predicting "job satisfaction" is of a different nature than that of predicting success, not only with respect to the types of test data that might prove most usable but also the techniques used in making predictions and the degree to which one could feel confidence in making predictions.

SUMMARY STATEMENT

Although we have listed illustrative problems facing a junior high school principal, the decisions involved are shared by teachers, counselors, and others engaged in the complex job of educating and guiding youth. Moreover, these problems are not confined to one age or maturity level. If a similar list of problems had been prepared by an *elementary* school principal, problems 1, 2, 3, and 5 might have been much the same; problem 4 would have been similar, although the areas in which national norms would have been stressed would be the "fundamental skills," while the committee of local teachers might have been needed to develop a test on a social studies area, such as Latin America. And since elementary school teachers are permitted wide latitude in choosing the content through which unit objectives are realized, a committee might decide against the construction of a city-wide test, preferring to develop a pool of items from which individual teachers could draw in developing their own classroom tests. Although problems 6 and 7 (as stated) are foreign to the elementary schools, similar prediction problems do occur at that level, for example, grouping children into reading and other instructional groups within the classroom on the basis of reading readiness scores and teacher judgment, as well as helping parents to decide whether their students are likely to benefit from special remedial services or from being accelerated or held back.

Carefully selected tests can provide data that will *aid* in making many professional judgments and decisions. However, a test that is best for checking on how well students have mastered the minimum essentials of a course may not be best for predicting how well they will achieve in a similar course of a higher level. We need to study each type of inference or decision made for which we will be using test results.

Obviously, in some of the problems listed above, sound decisions require the use of published tests for which *national norms* are available, that is, data regarding the test performance of representative groups of students in the nation's schools. When we are asked to inform the superintendent and community about students' knowledge of American history, we must admit that we have no absolute yardstick. But we can, if we use a test with national norms, interpret students' scores in terms of how they compare with the success of students of similar ability throughout the country.

On the other hand, for problem 3 (on the use of teaching machines in spelling instruction), national norms are *not* needed; and a local test would be more

adequate than a standardized one. We could take every twentieth word from the 500 words in the state speller to make a representative 25-word test. If the average score on this test at the end of the year is 20 words, or 80 percent, we can infer *with a fair degree of confidence* that students can spell 80 percent of the entire list of 500 words. Types of decisions for which we need norms, and types of local and national norms, are discussed in Chapter 2.

The expression, "with a fair degree of confidence," used in the preceding paragraph, refers to the concept of errors in measurement. Certainly we saved a good deal of time when we used a short 25-word test rather than one with one hundred words or more. But wouldn't one sample of only 25 words tend to favor some students who, by chance, had drilled on certain words that happened to be on that list? Since problem 3 involved a comparison of *averages*,³ we are on fairly safe ground in making the inference concerning the level of achievement on the total list. If we were going to use such a test for making inferences about the achievement of *individuals*, however, and for assigning individual grades in spelling, the test would need to be longer so as to minimize the effect of the chance factors involved in any sampling process. The sources of errors in measurement and methods of estimating the size of measurement errors will be considered in Chapter 3.

On some problems, such as the spelling study, we can easily devise a test that constitutes a *representative* sampling of the content in which we are interested. One or more random samplings of the state speller list of five hundred words are easily obtained. In developing the test of state history, we again wish to sample a universe⁴ of possible items, but it is impossible to define that universe as precisely as we could do with the spelling words. The question of how to achieve representativeness of content in such situations is considered in Chapter 4. We will also study predictive validity, which is involved whenever we use tests, ratings, or personal judgments to select students for a special class or for a scholarship, whenever we group students, or whenever we help students with decisions regarding subject or vocation choices. A test that might be quite valuable for assessing student learning of a representative sampling of spelling words (allowing us to make sound inferences concerning student proficiency on the total list) might have little value in *predicting* whether students should take an enriched English course or choose a career in secretarial work. These and other concepts related to the validity of tests and other measures will be studied in Chapter 4.

³ Obviously, a very large number of spelling tests could have been composed from all possible combinations of 25 words from the population of 500. When we learn how to determine, by methods described in Chapter 3, how well students' scores on one sample test agree with those on another sample, we can estimate the amount of error variance in students' scores and the standard error of measurement of a student's score. It is sufficient to emphasize here that the standard error of an *average* score for a large group of students is very small, in comparison with the standard error for individual scores. For example, in a problem of this type, if a group of 200-300 students averaged 80 percent spelling words correct, we could infer that the students knew 79-81 percent of the 500 words in the spelling list sampled.

⁴ Throughout the textbook, the term "sample" refers to a group of test items used or a group of individuals tested, while the term "universe" or "population" designates the larger defined group of which the sample is supposed to be representative.

In Chapter 5, we will illustrate how we can apply in the process of test selection the concepts developed in Chapters 2, 3, and 4, with respect to norms, reliability (errors in measurement), and validity. Chapter 5 will also consider the aids that the profession has developed to help us in this process of appraising published tests.

SELECTED REFERENCES

- ADKINS, DOROTHY C., "Measurement in Relation to the Educational Process," *Educational and Psychological Measurement*, vol. 18 (Summer 1958), pp. 221-240.
- BORDIN, EDWARD S., "Ethical Responsibilities of Instructors in Testing Courses," *Educational and Psychological Measurement*, vol. 11 (Autumn 1951), pp. 383-386.
- EBEL, ROBERT, AND DORA DAMRIN, "Tests and Examinations," in C. W. Harris, ed., *Encyclopedia of Educational Research*, 3d ed. New York: The Macmillan Company, 1960, pp. 1502-1517.
- SCATES, DOUGLAS E., "Some Problems Connected with Evaluation," *Journal of Educational Research*, vol. 45 (April 1952), pp. 599-608.
- SIEVERS, FRANK L., AND OTHERS, "Testing Issue," *School Life*, vol. 42 (September 1959), pp. 3-27.
- "Testing and Evaluation," *National Education Association Journal*, vol. 48 (November 1959), pp. 15-31.
- WRIGHTSTONE, J. WAYNE, *What Tests Can Tell Us about Children*. Chicago: Science Research Associates, 1954.

Interpreting Test Data in Terms of Converted Scores

Before we consider the more complex problems involved in test construction and test selection, let us consider the concepts involved in the interpretation of test data already collected.

Students' scores on a test (usually the number of items answered correctly) have little meaning except to indicate the relative position of each student in the class on each section of the test. Such untreated scores are known as *raw scores*. A student's raw score on one section of a test is not directly comparable to his raw score on another section, which may have a larger or smaller number of items of greater or less difficulty. Before such scores can be used to appraise a student's relative strengths and weaknesses, they must be expressed in comparable units. In other words, the raw scores must be translated into *converted scores*, which show (1) how a student's performance on the test compares with some arbitrary standard (such as a perfect test score) or (2) how his score compares with the scores of others in his class, his school, or some other group with whom he can appropriately be compared.

After the necessary test construction and test administration had been completed, Mr. Smith and his staff had available to them:

1. The results for six seventh-grade classes on the spelling test.
2. Results for all eighth-graders at Central Junior High School on a locally devised arithmetic test, developed by the eighth-grade teachers of that school.
3. The results for all eighth-grade students in the school district on both a locally developed state history test and a standardized United States history test.

For the standardized United States history test, the publishers had developed norms on students representative of the national population of

eighth-graders; hence each student's raw score could be interpreted in terms of percentile ranks (which will be explained in a later section). For the local tests, however, no such norms were available.

CONVERTED SCORES BASED ON COMPARISON WITH A PERFECT SCORE

The research committee on spelling had decided that there was no need for norms. Each student's score could be compared with a perfect score on the test. From a "percentage correct" score on the spelling test, one could infer the average percentage of seventh-grade spelling words that a pupil could spell. And this was the sole purpose of the spelling test.¹

The city-wide committee on history also intended to translate raw scores on the state history test into "percentage correct" scores. In fact, they had designed the final edition of the history test to have one hundred items, so that each student's raw score would be his "percentage correct" score.

"Percentage correct" scores, however, are most meaningful when we are able to define a universe of learnings (as we did in spelling) and sample it in such a way that the test is representative of this defined universe.² The ease with which the spelling committee had been able to take a random sampling of a defined universe of spelling words made the history committee members wish that all evaluation problems in education were as simple.

Students' "percentage correct" scores on the history test are comparable with those on another test, for example, the arithmetic test, only when two tests are equally difficult and when students' scores show equal "scatter" or variation around the average. Actually, tests rarely meet these conditions unless great care has been taken in their construction. Hence teachers are not justified in drawing inferences about improvements or retrogressions in student achievement when "percentage correct" scores increase or decrease throughout the school year.

"Percentage correct" scores provide no answers to such questions as faced Mr. Smith and his teachers:

¹ The student will recall that construction of the spelling test had only required the selection of every twentieth word from a defined universe of 500 seventh-grade spelling words and that from a student's "percentage-right" score, one could easily infer his probable level of achievement on the entire 500 words.

² Throughout the textbook, the term "sample" is used to refer to a group of test items used, or a group of individuals tested, while the term "population" or "universe" designates the larger defined group, of which the sample is supposed to be representative.

1. Was the average score in the school (or in a class group) as high as it should be?
2. On the average, was the school (or a class group) doing as well in their knowledge of state history as they were in United States history? And how did the students' achievements in both areas of history compare with their achievement in arithmetic?
3. Was a given individual doing as well in state history as he was in United States history? or in arithmetic?

Answering their first question inevitably involves professional judgment. However, certain procedures provide information helpful in making such professional judgments:

- a. comparing the achievement of their students with those of eighth-graders in the country as a whole (which could be done for the United States history test if the sample on which the test was standardized seemed sufficiently representative).³
- b. comparing student achievement with some external criterion, for example, checking to see how many students attained a level of arithmetic achievement associated with satisfactory performance in general mathematics or algebra courses.
- c. having representative teachers rate each test question as "essential," "desirable," and the like and then determining what percentage of students answered these most significant questions accurately. Here again, subjective judgment would be involved in deciding what level of achievement was satisfactory.

The second and third questions could best be answered by developing local norms. Such norms could be used to translate raw scores to converted scores, which would be comparable from test to test. The running of an adding machine tape had revealed the following average scores on the three tests. The term "mean" is used in statistical work for this type of average. $\text{Mean} = \frac{\Sigma X}{N}$ where Σ is a symbol meaning "the sum of"; X stands for score, and N , for number of cases.

		MEAN (AVERAGE) RAW SCORE	MEAN (AVERAGE) PERCENT CORRECT
Arithmetic test	(100 items)	84	84
State history test	(100 items)	78	78
United States history test	(175 items)	109	62

Obviously, if a hypothetical student had made an average "percent correct" score in each test, we could *not* infer that he had done best in arithmetic,

³ The factors that should be considered in attaining representativeness in norming samples are considered on pp. 60-62.

next best in state history, and least well in United States history. If a student happened to earn these scores on the three tests, we would have to say that he had done equally well on all three tests (if we used comparison with other students as our basis of interpretation). All the remaining types of converted scores, to be discussed in this chapter, involve such inter-individual comparisons.

CONVERTED SCORES BASED ON COMPARISONS AMONG EXAMINEES

There are three main approaches to obtaining converted scores, on the basis of comparisons among individuals. These approaches, as summarized in Table 2.7, include:

1. Comparison in terms of the difference between a student's score and the group average or mean,⁴ this difference to be expressed in terms of a standard unit (the standard deviation⁵ or some multiple thereof).
2. Comparison in terms of the rank of the student's score within the group of all students tested (or some defined reference group).
3. Comparison in terms of the average age or grade status of students obtaining the same score.

The third approach (age or grade norms) would be unsuitable for the state history test. It would be meaningless to interpret a student's score in eighth-grade history as average for a ninth-grader or tenth-grader. This history course is given only in eighth grade; hence, at the end of the eighth grade, students would earn a higher average score than they would if tested a year or two later at these higher grade levels.

Both the first and second types of norms, however, are suitable, and each type has certain advantages. Percentile scores (the second type) are more easily interpreted; but standard scores (the first type) have units of

⁴ The "mean" (M) is a term used to designate the type of average computed by totaling all scores and dividing the sum by the number of cases.

⁵ The term "variability" refers to the extent to which scores are clustered closely around the average or more widely dispersed. For example, the variability of all high school students with respect to height would be greater than the variability for either boys or girls, considered as separate groups. Of the frequently used measures of variability, the standard deviation (SD) is the most stable and meaningful. (J. P. Guilford, *Fundamental Statistics in Psychology and Education*, 3d ed. (New York: McGraw-Hill Book Company, Inc., 1956), p. 99. The SD can be computed for a set of data by the following formula: $SD = \sqrt{\frac{\sum x^2}{N}}$ where Σ stands for "sum of," x = the difference between each raw score and the average, and N = the number of cases. An approximation formula for the SD will be used in this chapter.

equal size (that is, representing equal ranges in raw scores) throughout the scale. Moreover, standard scores can be used in computing averages and making other needed computations, whereas one cannot average percentile scores.

Standard Scores

Standard scores, although more difficult to understand than percentile scores, have many advantages. In computing a *z-score*, which is the basic type of standard score, one finds the difference (or deviation) between a student's raw score (X) and the average or mean (M) for his school (or other reference group). Then this deviation (x) is divided by the standard deviation (which constitutes a standard unit of measurement). The formula may be written in either of two ways:

$$z = \frac{X - M}{SD} \text{ or}$$

$$z = \frac{x}{SD}$$

That is, we first find how much the student's score falls above or below the mean. Then, if we are going to compare a student's performance on this test with his performance on other tests, this difference must be expressed in terms of some standard unit.

The need for some standard unit is apparent when we recognize that a score ten points above the average on a 25-problem arithmetic test may be the highest score, while a score ten points above the average on a 300-word vocabulary test may represent an insignificant variation from the average. Not only is length of test important but the extent of dispersion or variability in student scores. If a test is very easy for a group (as a 300-word high school vocabulary test would be for college students), the dispersion in scores may be small even though the test is long. That is, almost all the students' scores may be clustered close together.

Test specialists have found that the best procedure for obtaining scores that are comparable from test to test is to divide each deviation score by a measure of the dispersion or variability of student scores around the average. The *SD* (standard deviation) has come to be preferred because of its meaningfulness and its broad applicability.

THE STANDARD DEVIATION AS A UNIT OF MEASUREMENT The standard deviation is used as the common unit of measurement in comparing test data and other educational measurements. Such a unit is greatly needed in education because of the impossibility of establishing a zero point or a

20.7.93

7095

maximum for such attributes as scholastic aptitude, competency in hand-writing, or social adjustment. Before we learn how to compute the standard deviation, we should examine the type of distribution frequently found when we graph the frequency (number of cases) for each score.

When a large number of individuals, selected at random, are measured with respect to almost any dimension (height, shoulder width, scholastic aptitude, and the like), a graph showing the frequency, or number of cases, resembles the normal, or probability, curve; that is, the frequency distribution is similar to the bell-shaped distribution shown in Figure 2.1. There tend to be relatively few cases with scores at each extreme and relatively large numbers of cases with scores near the average.

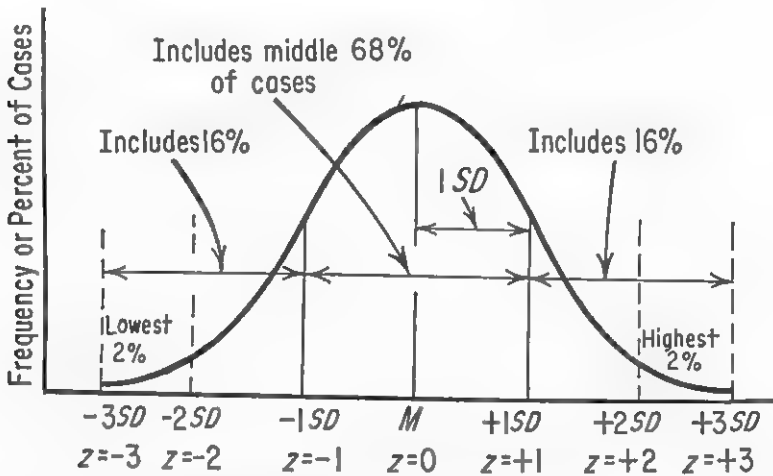


Fig. 2.1 A Normal Distribution Curve Showing Mean and Standard Deviation

In a perfectly normal distribution, as shown in Figure 2.1, the middle 68 percent of the cases are included between scores one *SD* below the *M* and one *SD* above the *M*. Once the size of the *SD* is determined, any score or other measure can be interpreted in terms of its deviation above or below the mean in units of standard deviation.

For example, if the heights of 1000 seventh-grade boys were measured and the frequency (number of cases) for each height were graphed, the result would approximate a bell-shaped curve, called the *normal curve*. Graphing the frequencies for each height for 1000 ninth-grade boys would result in another approximately normal curve with a somewhat higher mean height and somewhat greater variability. The heights of all seventh-grade boys could be translated into *z*-scores, using the seventh grade mean and *SD*; while those for ninth-grade boys could be converted into *z*-scores on the basis of the ninth-grade mean and *SD*. Thus, it is possible to com-

pare a certain boy's height with those of others his own age when he is in the seventh and later in the ninth grade. The comparison of such converted scores would reveal whether he maintained the same relative position in the two age groups or whether he was relatively shorter or taller in the ninth grade than he had been in the seventh grade. At each grade level the deviation of his height from the mean would be divided by the *SD* for that age group.

AN ILLUSTRATION OF COMPUTED Z-SCORES (BASIC STANDARD SCORES) Since all readers are familiar with IQ's, they will be used as our first illustration of the computation of *z*-scores. Here the mean (*M*) for the general population is 100, and the standard deviation (*SD*) is 15 to 16 IQ points. We will assume an *SD* of 15. Hence a student with an IQ of 115 has a *z*-score of 1.0 because he is one *SD* above the mean. Similarly, a student with an IQ of 130 has a *z*-score of 2.0.

$$z = \frac{X - M}{SD} \quad \text{When } X = 130 \quad z = \frac{130 - 100}{15} = +2.0$$

$$\text{When } X = 115 \quad z = \frac{115 - 100}{15} = +1.0$$

or

$$z = \frac{x}{D} \quad \text{When } X = 100 \quad z = \frac{100 - 100}{15} = 0.0$$

$$\text{When } X = 85 \quad z = \frac{85 - 100}{15} = -1.0$$

where $x = X - M$

$$\text{When } X = 70 \quad z = \frac{70 - 100}{15} = -2.0$$

Equal differences in *z*-scores correspond to equal differences in the raw scores on which they are based.

COMPUTATION OF Z-SCORES AND T-SCORES FOR THE LOCAL TESTS It is evident that if we knew the *SD*'s and means for the scores on the history, arithmetic, and other tests, we could readily compute *z*-scores for all students by this simple formula. Then these scores on the achievement tests would be comparable with each other and with those on an intelligence test.⁶

⁶ One could infer, for example, that a student with an IQ of 115 (*z*-score of 1.0) who had a *z*-score of approximately 1.0 on the history test or any of the other tests was working approximately at ability level. Such a comparison of *z*-scores (based on a sampling of the nation's population, as in the IQ test), with *z*-scores based on locally developed tests would be valid only if the local IQ distribution was such that the mean was approximately 100 and the *SD* approximately 15.

In the preliminary analysis of the data for Central High School, an approximation formula for the *SD* will be adequate. The use of this simplified formula gives us an *SD* of 10 for scores on the state history test, given in Table 2.1. The mean has already been computed as 78.0 by simply totalling the scores and dividing the sum by *N* (the number of cases) ($M = \frac{\sum X}{N}$). The *M* and *SD* are then used to compute illustrative *z*-scores, shown in Tables 2.3 and 2.4.

Table 2.1
Test Scores for Central High School Students in State History Test

RAW SCORE				RAW SCORE			
SCORE		FREQUENCY		SCORE		FREQUENCY	
<i>X</i>	TALLIES	<i>f</i>	<i>fX</i>	<i>X</i>	TALLIES	<i>f</i>	<i>fX</i>
98		2	196	75		4	
97		2	194	74		5	
96				73		4	
95		1	95	72		3	
94		2	188	71		3	
93		2	186	70		3	
92		2	184	69		5	
91		3	273	68		3	204
90		3	270	67		2	134
89		2	178	66		3	198
88		1	88	65		3	195
87		3		64		1	64
86		4		63		2	126
85		3		62		1	62
84		4		61			
83		4		60		1	60
82		4		59		2	118
81		5		58		1	58
80		5		57			
79		4		56			
78		7		55			
77		4		54		1	54
76		6					

Sum of
highest
1/6 of
scores
= 1852

Sum of
lowest
1/6 of
scores
= 1273

$$\text{Standard deviation}^a = \frac{\text{Sum of high sixth} - \text{sum of low sixth}}{\text{Half the number of students}}$$

$$SD = \frac{1852 - 1273}{60} = \frac{579}{60} = 9.65 = 10.$$

^a Formula taken from Paul B. Diederich *Short-Cut Statistics for Teacher-Made Tests*, Evaluation and Advisory Service Series No. 5 (Princeton, N. J.: Educational Testing Service, 1960), p. 21.

Table 2.2

Frequency Distribution for Central High School Students in State History Test and Computation Guide for Obtaining the Mean by the Short Method

SCORE INTERVAL ^a	<i>f</i>	<i>d</i>	<i>fd</i>	DIRECTIONS FOR COMPUTING MEAN FROM GROUPED DATA
96-98	3	+6	+18	1. Choose any interval as an <i>arbitrary origin</i> . Here the interval 78-80 has been chosen.
93-95	5	+5	+25	
90-92	8	+4	+32	2. Assign a <i>d</i> value to each interval (in terms of the number of intervals it lies above or below the arbitrary origin).
87-89	7	+3	+21	
84-86	11	+2	+22	3. Then, in each row, multiply the entries in the <i>f</i> and <i>d</i> columns, entering products in the <i>fd</i> column.
81-83	13	+1	+13	
78-80	16	0	(+131) ^b	4. Add the <i>fd</i> column to obtain Σfd (Σ is a symbol for "the sum of").
75-77	14	-1	-14	
72-74	12	-2	-24	5. Obtain the correction (in intervals) by dividing Σfd by <i>N</i> (number of cases).
69-71	11	-3	-33	
66-68	8	-4	-32	6. Multiply the correction by <i>i</i> (size of interval) to obtain the correction in score points. Then add it algebraically to the assumed mean (the midpoint of the interval selected as arbitrary origin).
63-65	6	-5	-30	
60-62	2	-6	-12	
57-59	3	-7	-21	
54-56	1	-8	-8	
<i>N</i> = 120			(-174) ^b	
		$\Sigma fd = -43$		

Assumed Mean (*AM*) = midpoint of 78-80 interval = 79

$$\text{Correction } (c) = \frac{\Sigma fd}{N} = \frac{-43}{120} = -0.36$$

$$\begin{aligned} \text{Mean } (M) &= AM + (c)(i) \\ &= 79.00 + (-0.36)(3) = 79.00 - 1.08 = \\ &= 77.92 = 78 \end{aligned}$$

^a In order to estimate the size of interval (*i*) to be used for tallying test scores, one can divide the range of scores by 15. In this case, the range of scores was from 54 to 98, or 44 points. The size of interval was $\frac{44}{15}$, or 3. Division by 15 is recommended since at least 15 intervals are desirable. A smaller number of intervals usually involves too much loss of information about the distribution of scores and increases the error in computation that results from not using precise score values. Ordinarily, intervals of 3, 5, 10, or multiples of 10 make for ease in tallying.

^b Partial sums of positive and negative *fd* values.

Table 2.3
Frequency Distribution for Central High School Students
in an Arithmetic Test

SCORE INTERVAL	<i>f</i>	<i>d</i>	<i>fd</i>	
96-98	21	+5	+105	Examples of computation of standard scores for skewed distribution
93-95	18	+4	+ 72	
90-92	12	+3	+ 36	
87-89	10	+2	+ 20	
84-86	8	+1	+ 8	$z = \frac{\text{Raw Score} - \text{Mean}}{\text{Standard deviation}}$
81-83	7	0	(+241)	$M = 84 \quad SD = 11$
78-80	8	-1	- 8	
75-77	6	-2	- 12	Highest score
72-74	8	-3	- 24	$X = 98 \quad z = \frac{98 - 84}{11} = \frac{14}{11} = 1.3$
69-71	5	-4	- 20	An average score
66-68	4	-5	- 20	$X = 84 \quad z = \frac{84 - 84}{11} = \frac{0}{11} = 0$
63-65	4	-6	- 24	
60-62	3	-7	- 21	Lowest score
57-59	3	-8	- 24	$X = 54 \quad z = \frac{54 - 84}{11} = \frac{-30}{11} = -2.7$
54-56	3	-9	- 27	
	120		(-180)	
		$\Sigma fd = + 61$		Other examples
				$X = 90 \quad z = \frac{90 - 84}{11} = \frac{6}{11} = .5$
				$X = 80 \quad z = \frac{80 - 84}{11} = \frac{-4}{11} = -.4$
				$X = 70 \quad z = \frac{70 - 84}{11} = \frac{-14}{11} = -1.3$
				$X = 60 \quad z = \frac{60 - 84}{11} = \frac{-24}{11} = -2.2$

Assumed Mean (*AM*) = midpoint of 81-83 interval = 82

$$\text{Correction } (c) = \frac{\Sigma fd}{N} = \frac{+61}{120} = +.51$$

$$M = AM + (c)(i) = 82 + (.51)(3) = 83.53 = 84$$

$$\text{Stand. Dev. } (SD) = \frac{\text{Sum of high sixth} - \text{sum of low sixth}}{\text{Half the number of students}} =$$

$$\frac{1943 - 1260}{60} = \frac{683}{60} = 11.4$$

The distribution of history scores in Table 2.2 is highly symmetrical and approaches a normal curve, while the distribution of arithmetic scores in Table 2.3 is skewed, or not symmetrical. Students' scores are piled up at the high-score end of the distribution, with more than half the scores in the top four intervals. In both the normal and the skewed distribution, differences in *z*-scores faithfully reflect proportional differences in raw scores, as shown in Table 2.4.

Table 2.4
Comparison of *z*-scores and *T*-scores for Students with Identical
Raw Scores on Arithmetic and State History Tests

	Raw scores ^a		<i>z</i> -scores ^b		<i>T</i> -scores ^b	
	ARITH.	STATE HISTORY	ARITH.	STATE HISTORY	ARITH.	STATE HISTORY
James	90	90	0.5	1.2	55	62
Mary	80	80	-0.4	+0.2	46	52
Sandra	70	70	-1.3	-0.8	37	42
John	60	60	-2.2	-1.8	28	32

^a Note that although both these tests are 100-item tests, identical raw scores on the two tests are translated into different *z*-scores and *T*-scores because of differences in the difficulty of the tests and slight differences in the *SD* values, which reflect the degree to which scores are dispersed around the mean. For example, a raw score of 90 in the state history test represents greater superiority in comparison with other students than does a score of 90 on the arithmetic test.

^b Equal differences in raw scores are accurately reflected in equal differences in standard scores, that is, Mary and James differ by 10 raw score points in arithmetic; so also do Sandra and John. The difference in their *z*-scores in arithmetic is 0.9 for each pair; the difference in their *T*-scores is 9 points for each pair. The formula for the *T*-score is as follows: $T = 10z + 50$.

The *z*-scores, although easy to compute, have the disadvantage of involving decimal points and minus signs. Hence, it may be desirable to translate *z*-scores into equivalent *T*-scores,⁷ which avoid these two problems. For *T*-scores, the mean on any test is equated to 50 and the *SD*, to 10.

Since we will later be working with large numbers of cases, the students may wish to learn a method of computing the mean from grouped data. In Table 2.2, the short method of computing the mean from grouped data is illustrated for the history test. In each case, the test data have been

⁷ $T\text{-score} = 10z + 50$.

tallied by an interval of 3. This method gives almost identical results⁸ to those obtained by adding original scores and dividing the total by the number of cases.

Converted Scores Based on Student's Rank within Group

We will now consider types of converted scores, based on the student's rank within a group. It was obviously of limited value for the city-wide history committee to compute "rank in class" for each student, since classes varied in size, or to compute "rank in school," since the size of eighth-grade classes varied from 80 to 170 in the different junior high schools. To rank fortieth in the smallest school would be to rank close to the average; to rank fortieth in the largest school would be to be in the top one-fourth of the group.

PERCENTILE SCORES A conference with the research director convinced the history committee that the use of percentile scores was advisable. A simple graphic procedure would make it possible to find the percentile scores corresponding to each raw score. A student with a median score would have a percentile score of 50, since his raw score exceeded those of 50 percent of the group; a student would have a percentile score of 90 if his raw score exceeded those for 90 percent of the group.

Mr. Smith decided to experiment with these procedures with the data for Central High School before he applied them to the city-wide results. He computed the cumulative frequency and cumulative percentage for each interval, as shown in Table 2.5. He then used a formula to compute the score value for P_{40} , the score that exceeded those for 40 percent of the group. The work involved to obtain only one of the 99 percentile scores by formula convinced him that it was best to shift to the graphic method.

In Figure 2.2, the cumulative percentage for each score interval is plotted in a curve, called the *ogive*. From this graph, the percentile score corresponding to each raw score can be read. The value read from the graph for P_{40} is similar to the computed percentile score of 75.6. In this graph and its footnote, the procedures for obtaining a percentile score for

⁸ Means obtained by the usual method and by the short method (based on grouped data) would be approximately the same. The loss of accuracy by grouping data in score intervals, and treating all scores as if they fell at the midpoint of the interval, does result in a small "grouping error." In this case the difference is only 0.1, the mean computed by $\frac{\sum X}{N}$ being 78.0, and the mean computed by the short method in Table 2.2 being 77.9. This small error is introduced by *grouping* the data, rather than by using the "short method," which simply reduces the amount of computation.

Table 2.5
Cumulative Frequencies and Cumulative Percentages Used in Graphing
Ogives for State History Test and Local Arithmetic Test

SCORE INTERVALS	State history test			Arithmetic test		
	FRE- QUENCY	CUMU- LATIVE FREQUENCY f_c	CUMU- LATIVE PERCENT	FRE- QUENCY	CUMU- LATIVE FREQUENCY f_c	CUMU- LATIVE PERCENT
96-98	3	120	100.0	21	120	100
93-95	5	117	97.4	18	99	83
90-92	8	112	93.3	12	81	68
87-89	7	104	86.6	10	69	58
84-86	11	97	80.8	8	59	49
81-83	13	86	71.7	7	51	43
78-80	16	73	60.8	8	44	37
75-77	14	57	47.5	6	36	30
72-74	12	43	35.8	8	30	25
69-71	11	31	25.8	5	22	18
66-68	8	20	16.7	4	17	14
63-65	6	12	10.0	4	13	11
60-62	2	6	5.0	3	9	7
57-59	3	4	3.3	3	6	5
54-56	1	1	0.8	3	3	2
	120			120		

Directions for Computing Percentile Points

The percentile point, below which falls any specified percentage of scores, may be obtained either by use of a graph (as in Figures 2.2 or 2.6) or by use of the following formula:

$$P = LL + \left(\frac{FN - f_c}{f_w} \right) \times i$$

Where LL = lower limit of interval containing the desired percentile point (always $\frac{1}{2}$ unit below the score limit of interval)

F = a fraction that varies with the percentile desired; for the Mdn or P_{50} (the point below which 50 percent of the cases fall), F is $\frac{50}{100}$; for P_{40} , F is $\frac{40}{100}$, and the like.

N = number of cases (or test scores)

f_c = cumulative frequency below the interval containing the desired percentile point

f_w = number of cases (or test scores) within the interval containing the desired percentile point

i = size of interval (in this distribution $i = 3$)

Table 2.5 (Continued)

Cumulative Frequencies and Cumulative Percentages Used in Graphing
Ogives for State History Test and Local Arithmetic Test

EXAMPLE If we are computing P_{40} , we first find FN , which is $\frac{40}{100} \times 120 = 48$.

Note that in the f_c column for the state history test, there are 43 cases below the interval (75-77) that contain P_{40} ; therefore $f_c = 43$ and $f_o = 14$

$$P_{40} = 74.5 + 3 \left(\frac{48 - 43}{14} \right) = 74.5 + 3 \left(\frac{5}{14} \right) = 74.5 + 1.1 = 75.6$$

a given raw score ($X = 70$), and for obtaining the raw score for a specified percentile score (P_{75}) are illustrated.

Percentile scores have many advantages. They are easily obtained by the graphic method. They can be easily interpreted to students and parents, without bringing in the concept of standard deviation, which is more difficult to understand. Percentile scores, however, have certain disadvantages. As a measure of growth, they can be misleading; for example, a student who has a percentile score of 60 at the beginning of the seventh grade, and also at the beginning of the eighth grade, has made normal progress. In each case, he is being compared with students of his own grade level; and he has maintained the same relative status within his grade level group.

Another disadvantage of percentile scores can be noted from an examination of Figure 2.2. A difference of 20 percentile points near the middle of the distribution represents a relatively small difference in raw scores, whereas a difference of 20 percentile points near either extreme represents a much larger difference in raw scores.⁹ This characteristic of percentile scores must be taken into account in interpreting test data; or one would tend to overestimate the significance of differences in percentile scores near the average, and underestimate those near the extremes. A difference of ten percentile points between P_{45} and P_{55} (both indicating achievement near the average of the group) represents a much smaller variation in achievement than the difference between P_{10} and P_{20} , or between P_{80} and P_{90} . To take height as an example, it is readily apparent that if a group of 100 men were arranged in order of height, the middle 10 percent of the group (corresponding to percentile scores between 45 and 55) would be almost identical in height. However, at either extreme, one would find marked differences in height corresponding to a difference of ten percentile points. For example, the tallest man would be conspicuously taller

⁹ For example, in Figure 2.2, the difference of 20 percentile points between P_{40} and P_{60} represents a range of only 6 raw scores from 75 to 81; while a difference of 20 percentile points from P_{70} to P_{90} represents a difference of 12 raw score points, that is, from 86 to 98.

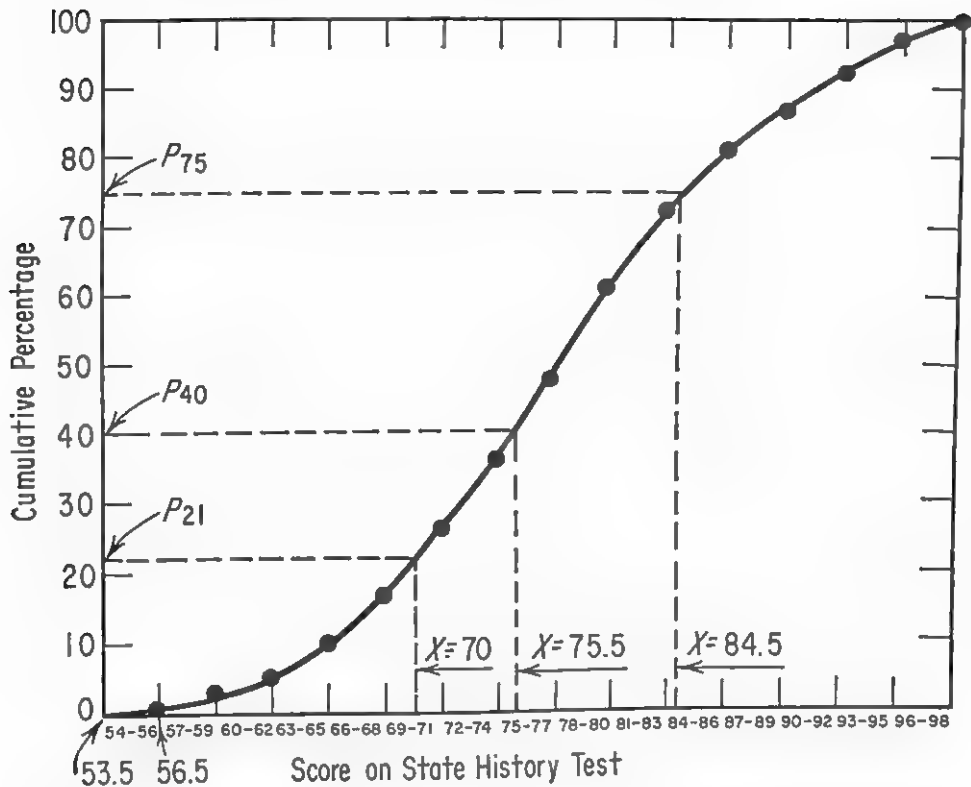


Fig. 2.2 Ogive, or Cumulated Percentage Curve, for Scores on State History Test (based on cumulative percent column in Table 2.5)

NOTE: To estimate the percentile equivalent of a raw score, take a card, position the right edge of the card at the raw-score value and perpendicular to the base line of the graph. Move the card up or down until the horizontal edge intersects the curve. With the card properly positioned, read the desired percentile point on the percentile scale. In this case, with the right edge perpendicular to the base line at $X = 70$, the card would intersect with the curve at a point corresponding to a percentile point of 21. To find the score value corresponding to any percentile point, position the upper edge of the card at the percentage value desired. Move the card over until the right edge intersects the curve. Then with the card properly positioned, follow down to the base line and read off the score value. In this case P_{75} corresponds to a score value of approximately 84.5. Note that the value of the cumulative percentage is plotted at the upper limit of each score interval and that the lower and upper limits of each interval are respectively at $\frac{1}{2}$ unit below and above the score points shown.

than the man ranking at the 90th percentile, and the shortest would be markedly shorter than the man at the 10th percentile. For further clarification of this concept, see Figure A.1 in Appendix D and the accompanying explanation.

Mr. Smith also realized that he could not average percentile scores directly. To obtain an average percentile score for two or more individuals, one must first obtain the raw score equivalents of the percentile scores, average them, and find the equivalent percentile score for the average raw score.

Up to this point, we have used the term "percentile score," but the student will find many test manuals using the term "percentile rank." The term "percentile rank" has a somewhat different meaning than the term "percentile point" computed in Table 2.5. The directions given in that table enable one to compute each of 99 percentile points, which divide the frequency distribution into one hundred groups containing equal numbers of cases. The "percentile rank" is a converted score for a *span* of raw score values, centering around the percentile point; a *PR* of 2 is the converted score for raw score values centered around P_2 . The following chart illustrates the difference between the two terms and also shows why the terms 99+ or 1- appear in norms tables of test manuals.

PERCENTILE POINTS	P_1	P_2	P_3	P_{97}	P_{98}	P_{99}	
	Lowest 1% of cases	Next lowest 1% of cases	Next lowest 1% of cases		Next highest 1% of cases	Next highest 1% of cases	Highest 1% of cases
PERCENTILE RANKS	1 —	1	2		98	99	99 +

Graphic explanation of percentile points and percentile ranks at the upper and lower extremes of a frequency distribution

When we are dealing with the division of a range of raw scores into *one hundred* parts, the distinction between percentile points and percentile ranks is a minor distinction. However, when we are dealing with the division of a distribution into *tenths*, the distinction becomes very important. For example, the raw scores that would correspond to a decile rank of 1 for the state history test would be scores that would include 5 percent of the cases on both sides of P_{10} or D_4 . Although decile ranks are infrequently used, the example is included to help the student distinguish between these two concepts.¹⁰

¹⁰ In Table 2.5, $P_{10} = 65.5$. Decile 1 would include scores from 63 to 68. See Howard B. Lyman, *Test Scores and What They Mean* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1963), p. 118.

DECILE POINTS		D_1	D_2	D_3	D_7	D_8	D_9	
DECILE RANKS	0	1	2			8	9	10

NORMALIZED STANDARD SCORES In an earlier section of this chapter, we used the *SD* as a unit for expressing test scores in a common frame of reference, which compensates for differences in variability between distributions of test scores. We have shown how to compute *z*-scores and *T*-scores so that student scores on different tests are expressed in comparable terms, that is, in terms of how much these scores differ from the mean (in *SD* units). We have obtained percentile equivalents of raw scores by computation and through the graphic method.

We would now like to show how percentile ranks and *normalized standard scores* can easily be obtained through the use of the Otis Normal Percentile Chart, if the distribution of test scores approaches normality. First, it is desirable to examine some of the characteristics of the "normal curve." (See Fig. 2.1.)

1. *Characteristics of the Normal Curve.* The normal probability curve or normal curve is a theoretical curve. However, the frequency curves for many attributes (which are affected by a multitude of interacting factors) approximate the normal curve. For example, many biological characteristics, such as height, width of shoulders, and the like are normally distributed, probably because of the myriad possibilities of different chance combinations of genes affecting these characteristics. Also many distributions of test scores approximate a normal distribution, especially when a large number of persons have been administered a test designed at an optimum level of difficulty.¹¹ And, as we shall see in the next chapter, a large number of repeated measurements of any characteristic, such as an individual's scores on repeated test samples of vocabulary, tend to be normally distributed, because of chance combinations of different measurement errors.

The fact that the normal curve is a theoretical curve that can be exactly described once the mean and *SD* are known, makes it exceedingly useful in the development of comparable scores for many diverse variables. For example, if such comparisons were useful, one could compute comparable standard scores on such diverse variables as height, speed of running, knowledge of vocabulary, and attitudes toward school. That is, one can define each individual's place on a normal distribution for a specified population and determine how much more or less he differs from the average in each variable.

¹¹ Optimum level of difficulty is discussed further in Chapter 5. For most purposes, the best level of difficulty is one which allows for maximum differentiation among the persons tested. This maximum differentiation is achieved when the average "number correct" is halfway between a chance score (for example, one half the number of items in a true-false or other two-choice test) and the maximum possible score.

Since the normal curve is so significant, we should note some of its major distinguishing characteristics:

- a. The largest number of cases are clustered in the center of the range, with the highest frequency at the exact center; thus the modal (most frequent) score is the same as the mean score, and also the same as the median (or midscore). A perpendicular line erected at the mid-point divides the area under the curve (and the number of cases it represents) in half.
- b. The curve is symmetrical, each half of the curve being the mirror image of the other.
- c. The shape of each half of the curve changes from convex to concave at a point 1 *SD* above and below the mean. About 68 percent of the area (and the cases) lie within one *SD* (plus and minus) of the mean.
- d. Tables exist that give much valuable information about the characteristics of the normal curve. That is, once we make the assumption that our data are approximately normally distributed, we can make a number of helpful inferences. For example, there is a definite relationship between (1) any *z*-score, and (2) the proportion of the area (or cases) which that *z*-score exceeds.

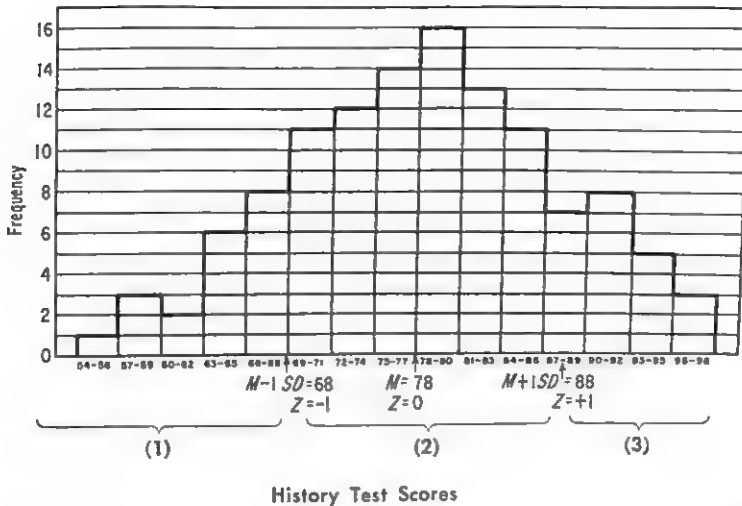
As an aid in conceptualizing the way in which "area under the normal curve" represents number, or proportion, of cases, the student is referred to Figure 2.3. In this histogram,¹² each case is represented by a small rectangle; hence it is apparent that the proportion of area represents the proportion of cases. The reader will note that since the distribution of history scores is approximately normal, the proportion of area (and cases) to the left of -1 *SD* (or to the right of $+1$ *SD*) is approximately the same as that found in a normal distribution.

From the left-hand side of Table A.1 in Appendix D we find that in a normal distribution a *z*-score of 0 exceeds 50 percent of the area (or cases); a *z*-score of 1 exceeds 84 percent; a *z*-score of 2 exceeds 98 percent; and the like. In the right half of the table, we find that a *z*-score of -1 exceeds 16 percent of the cases and a *z*-score of -2 exceeds 2 percent of the cases. Any other intermediate values for *z*-scores can easily be changed into percentile scores and vice versa. Examination of Figure 2.3 will show why these *z* values are equal to these percentile scores.

With Figure 2.4, on the other hand, in which the frequency distribution of scores is not symmetrical, but is skewed or asymmetrical, the percentages of cases in these same segments of the frequency polygon¹³ (below -1 *SD* and above $+1$ *SD*) do not approximate as closely the percentages in the normal curve.

¹² In a histogram, there is erected at each score interval a vertical bar representing the number of cases in that interval.

¹³ A frequency polygon is a simpler form than the histogram for showing the form of a frequency distribution. In the frequency polygon, the frequency for each interval is represented by a point plotted over the midpoint of that interval. The polygon is completed by connecting adjacent points with straight lines. At each extreme, the frequency curve is dropped to the baseline at the midpoint of the interval just above (or below) that containing the highest (or lowest) scores respectively.



- (1) 19 cases^a, or 16 percent of cases and 16 percent of area, are below 68, or a z-score of -1 in this distribution
 16 percent of the cases and 16 percent of the area are below -1 SD in a normal distribution
- (2) 81.5 cases, or 68 percent of cases and 68 percent of area, are between 68 and 88, or between z-scores of $+1$ and -1 in this distribution
 68 percent of the cases and 68 percent of the area are between $+1$ SD and -1 SD in a normal distribution
- (3) 19.5 cases^b, or 16 percent of cases and 16 percent of area, are above 88 or a z-score of $+1$ in this distribution
 16 percent of the cases and 16 percent of the area are above $+1$ SD in a normal distribution

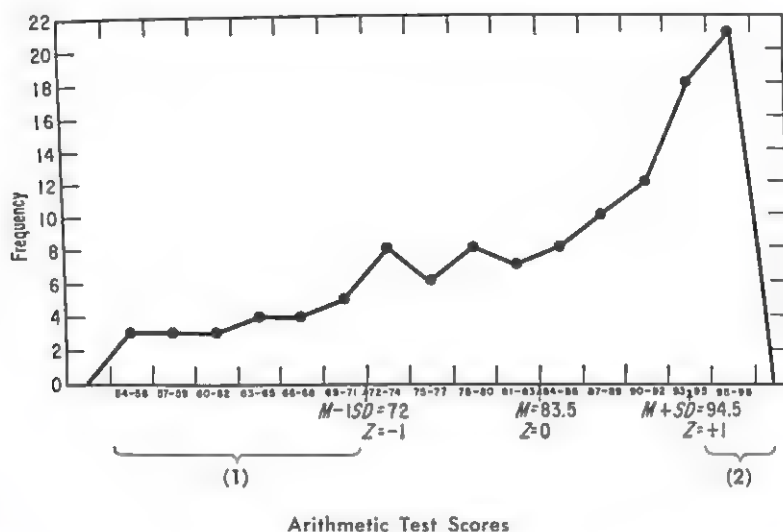
Fig. 2.3 Histogram for Distribution of Central High School Scores in the State History Test (designed to illustrate the concept that any specified area under a frequency curve represents the frequency, or percentage of cases, within its score boundaries).

^a Since 68 is 1/6 interval below the upper limit of the interval 66–68 (or 65.5–68.5), only 5/6 of the cases in the interval were included.

^b Since 88 is the midpoint of the 87–89 interval, one half of the cases in the 87–89 interval were included.

2. **Linear and Area Transformations of Test Scores.** The z-scores that we computed by formula, and the T-scores derived from them, are *linear* standard scores obtained through linear transformations (multiplying the original scores by a constant, and/or adding a constant to them). That is, the raw scores are translated by a formula to converted scores that preserve the original relationships between the raw scores. Equal differences in converted scores are proportional to equal differences in raw scores. The information

about differences in performance between students is preserved. If raw scores are plotted against their converted z-scores or T-scores, the points representing the pairs of raw and converted scores fall along a straight line (or show a *linear relationship*) (Figure 2.5).



(1) 23 cases^a, or 19 percent of the cases, are below 72, or a z-score of -1 in this distribution

16 percent of the cases are below a z-score of -1 in a normal distribution

(2) 27 cases^b, or 22.5 percent of the cases, are above 94.5, or a z-score of $+1$ in this distribution

16 percent of the cases are above a z-score of $+1$ in a normal distribution

Fig. 2.4 Frequency Polygon for Distribution of Central High School Scores in Arithmetic Test (designed to show skewed distribution obtained when test is too easy to discriminate adequately among high-achieving students).

^a Since 72 is $1/6$ interval above the lower limit of the interval 72-74 (71.5-74.5), $1/6$ of the cases in this interval were included.

^b Since 94.5 is $1/3$ interval below the upper limit of the interval 93-95 (92.5-95.5), $1/3$ of the cases in this interval were included.

In converting raw scores to *normalized standard scores*, however, the procedure is quite different. Raw scores are first converted into PR's (representing the proportion of cases or area exceeded); and these PR's, in turn, are converted into equivalent standard scores in a *normal distribution*. For example, in the following table we show the PR's for three raw scores on the state history test.

RAW SCORE	PERCENTILE RANK FROM FIGURE 2.2	NORMALIZED Z-SCORES (OBTAINED FROM TABLE A.1)	LINEAR Z-SCORES (OBTAINED BY FORMULA) TRANSFORMATION
70	21	-0.8	-0.8
76	40	-0.25	-0.2
85	75	+0.7	+0.7

If we use Table A.1 (based on percentile ranks or areas under the normal curve corresponding to different z-score values), we obtain the normalized standard scores shown in the third column. If we obtain z-scores by means of the formula $\left(z = \frac{X - M}{SD} \right)$, they are almost identical (as shown in the last column). The differences are negligible in this case because the frequency distribution of scores on the local state history test closely approximates a normal distribution.

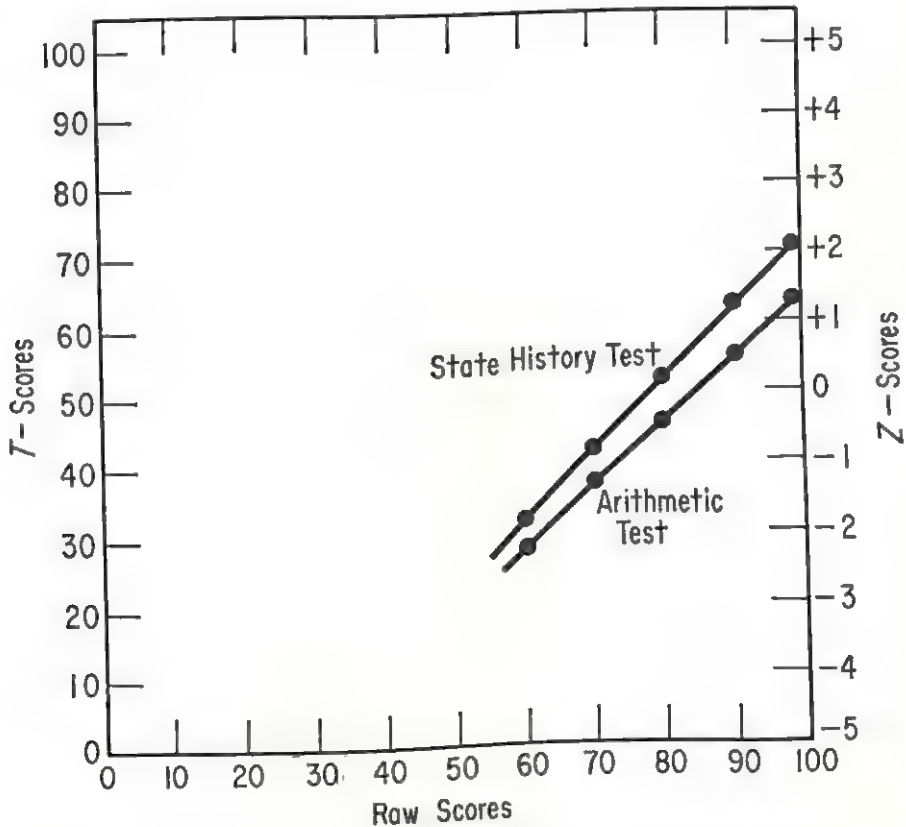


Fig. 2.5 Illustration of Linear (straight-line) Relationship between Raw Scores and Their Linear Transformations (z-scores and T-scores). State History Test Data from Table 2.6.

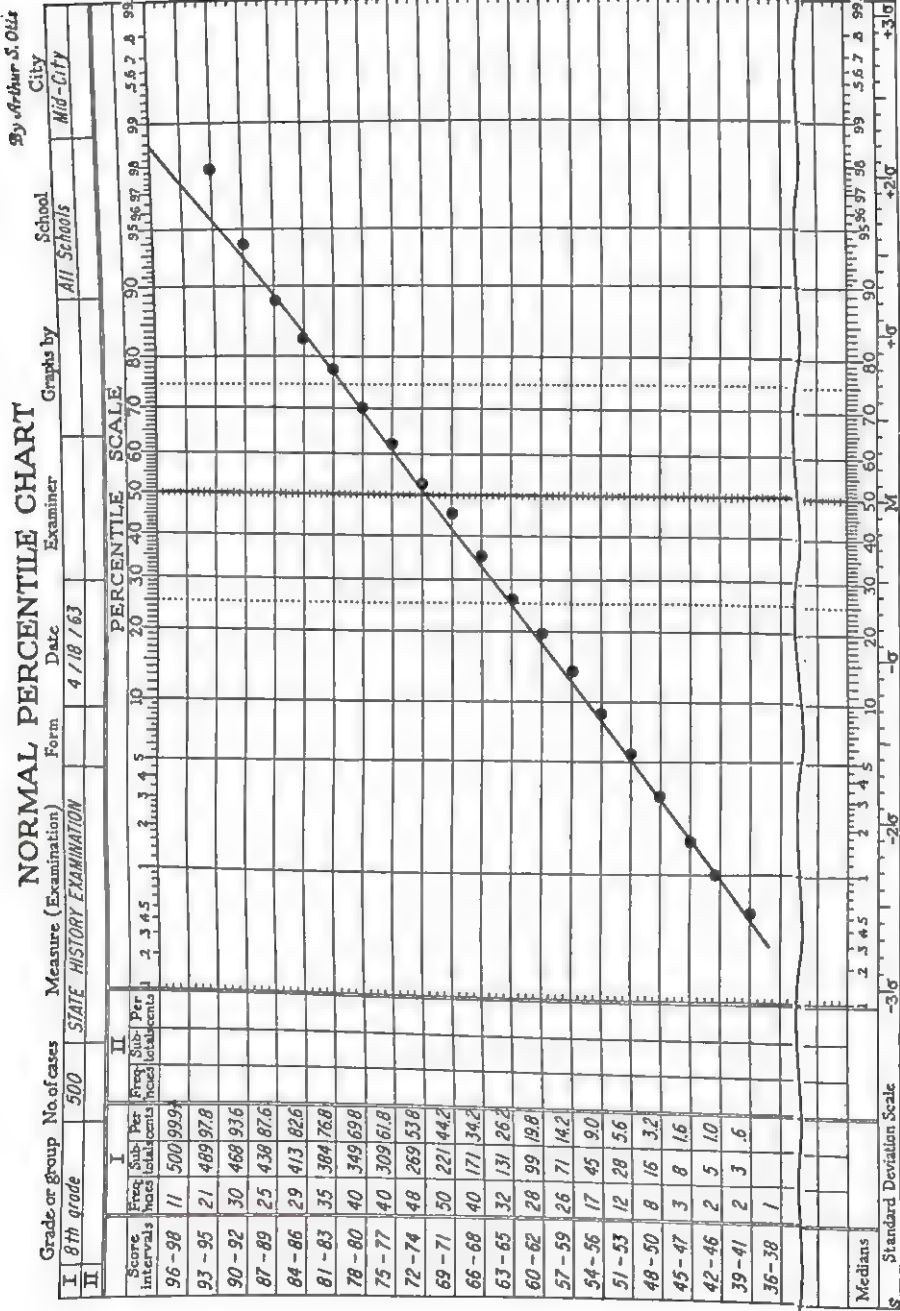


Fig. 2.6 Use of the Normal Percentile Chart in Obtaining Normalized Standard Scores—Data for All Schools. Normal Percentile Chart by Arthur S. Otis. Copyright 1938 by Harcourt, Brace & World, Inc., New York, N. Y. Copyright in Great Britain. All rights reserved. Reproduced by special permission.

Table 2.6
Percentile Ranks, *T*-scores, and Stanine Scores for the State History Test
for All Schools (as read from the Normal Percentile Chart,^a Figure 2.6)

SCORE	PERCENTILE RANK	<i>T</i> - SCORE	STANINE SCORE	SCORE	PERCENTILE RANK	<i>T</i> - SCORE	STANINE SCORE
45 & below	1	27	1	72	43	48	5
46-48	2	30		73	47	49	
49-50	3	31		74	50	50	
51-52	4	33		75	54	51	
53	5	34	2	76	56	52	6
54	6	35		77	59	52	
55	7	35		78	62	53	
56	8	36		79	65	54	
57	9	37	3	80	68	55	7
58	11	37		81	71	56	
59	12	38		82	74	57	
60	14	39		83	77	58	
61	16	40	4	84	79	58	8
62	18	41		85	81	59	
63	20	42		86	83	60	
64	22	43		87	85	61	
65	24	43	5	88	87	62	9
66	27	44		89	88	62	
67	30	45		90	90	63	
68	32	46		91	91	64	
69	34	46	6	92	92	64	9
70	37	47		93	93	65	
71	40	48		94	94	66	
				95	95	66	
				96	96	68	
				97 & above	99	73	

^a If a normal percentile chart is not available, normal probability paper can be used to obtain *T*-scores from a graph (with raw scores being plotted on the horizontal axis and cumulative proportions or percentages on the vertical axis). See J. P. Guilford, *Fundamental Statistics in Psychology*, third ed. (New York: McGraw-Hill Book Company, Inc., 1956), p. 498.

Normalized standard scores are obtained by a process known as "area transformation." In other words, a percentile score (representing a cumulative percentage or area) is used as the basis for obtaining a standard score. When area transformations are used, we are forcing our distribution of test scores to fit the shape of a normal distribution. If the frequency distribution of raw scores is approximately normal, normalized *T*-scores, or *T*-scaled scores,¹⁴ will be similar to *T*-scores obtained by formula.

¹⁴ Use of the term "*T*-scaled score," is recommended by Lyman to identify the normalized *T*-score, based on area transformation, as distinguished from the *T*-score based on linear transformation. Howard B. Lyman, *Test Scores and What They Mean* (Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1963), p. 115.

The following quotation from Anastasi indicates not only the conditions under which normalized standard scores may be used but also the desirability of adjusting the difficulty of test items so that the distribution of test scores more nearly approaches normality.

Normal standard scores are standard scores expressed in terms of a distribution that has been transformed to fit a normal curve. . . . Such a transformation should be carried out only when the sample is large and representative and when there is reason to believe that the deviation from normality results from defects in the test rather than from characteristics of the sample or from other factors affecting the behavior under consideration. . . . *Whenever feasible, it is generally more desirable to obtain a normal distribution of raw scores by proper adjustment of the difficulty level of test items, rather than by subsequently normalizing a markedly non-normal distribution.* [italics added]¹⁵

For example, if the arithmetic test were revised to include more difficult questions, the individual differences among students whose scores are now piled up at the high-score level would be more adequately measured. Then a more nearly symmetrical distribution of scores would be obtained, and the more efficient process of obtaining *PR*'s, *T*-scaled scores, and other normalized standard scores by means of Table A.1 or by means of the Otis Normal Percentile Chart could unquestionably be used.

We can easily obtain *PR*'s and the equivalent normalized standard scores if we graph the cumulated percentages on an Otis Normal Percentile Chart (Figure 2.6). In other words, a table of local norms for our state history test can be easily obtained by graphing the cumulated percentages for each interval and simply reading off the desired *PR*'s (top scale) or *z*-scores (bottom scale). A *T*-score scale or other types of normalized standard score scales could easily be added to the graph. On this kind of chart, intermediate values can easily be obtained, since a normal distribution appears on such a chart as a straight line. The values read from the graph for the state history test are shown in Table 2.6.

The graph was based on the data for all 500 students in the school district, shown in Table A.3, rather than those for Central High School. The reader will note by comparing Table A.3 with Table 2.2 that the mean score for the school district is lower than that for Central High School; and as one would expect, the *SD* for the school district is substantially larger than that for a single school (which would naturally be more homogeneous than "combined schools" with respect to achievement on any test). Table A.3 also provides a computing guide for obtaining the *SD* by the short method.

¹⁵ Anne Anastasi, *Psychological Testing* (New York: The Macmillan Company, 1954), p. 94.

STANINE SCORES For many purposes, a simpler type of one-digit standard score, called the stanine score,¹⁶ is desirable. In the stanine scale, raw scores are converted to a nine-point scale, with a mean of 5 and an *SD* of 2. These stanines represent approximately equal steps on the raw-score scale and can be used with any data that can be arranged in rank order. Stanine scores obtained for one test are comparable¹⁷ with those for any other test administered to the same group of students. Stanine scores can be averaged and used in other types of mathematical computation.

Stanine scores can be assigned to groups of raw scores by using the Otis Normal Percentile Chart, or simply by assigning stanine scores sequentially to raw scores, which have been ranked or tallied. The following distribution is used:

STANINE SCORES	1	2	3	4	5	6	7	8	9
PERCENT AT EACH LEVEL	Lowest 4%	Next lowest 7%	Next lowest 12%	Next lowest 17%	Middle 20%	Next highest 17%	Next highest 12%	Next highest 7%	Highest 4%

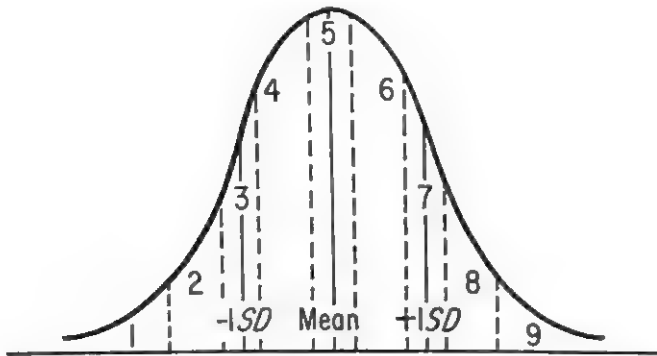
These percentages represent the areas included in defined segments of the normal curve, $\frac{1}{2}$ *SD* in width, as shown in Figure 2.7. If a teacher is assigning stanines to a distribution of test scores, he will not be able to follow these percentages exactly, assigning a score of 9 to the highest 4 percent, a score of 8 to the next 7 percent, and the like. Since one must obviously assign the same stanine score to all students receiving the same raw score, only an approximation of these percentages is possible.

Diederich¹⁸ suggests that teachers, in using stanines for their day-by-day equating of scores for recording in rollbooks, employ the following dis-

¹⁶ The term "stanine" is used because it is a STANDARD NINE-point scale from 9 (high) to 1 (low). One publisher to date (Harcourt, Brace & World) is using stanine norms with several of its standardized test batteries and has prepared a leaflet (free on request) on the use of stanines: Walter N. Durost, "The Characteristics, Use, and Computation of Stanines," *Test Service Notebook*, No. 23 (New York: Test Department, Harcourt, Brace & World, Inc., 1961). Many applications of the stanine technique are also explained in Walter N. Durost and George A. Prescott, *Essentials of Measurement for Teachers* (New York: Harcourt, Brace & World, Inc., 1962).

¹⁷ When we say that scores are "comparable," we are not implying that they are equivalent and can be substituted for each other. Two scores may be comparable and yet reflect quite different abilities. Two scores are comparable if they represent the same standing, or rank, in the same population.

¹⁸ Paul B. Diederich, *Short-cut Statistics for Teacher-made Tests*, Evaluation and Advisory Service Series No. 5 (Princeton, N. J.: Educational Testing Service, 1960, p. 37).



STANINE	RANGE (IN SD UNITS FROM MEAN)	IQ RANGE	PERCENTAGE AT EACH LEVEL	RANGE OF PERCENTILE RANKS
9	+1.75 SD units & above	128 & above	4	97 & above
8	+1.25 SD to +1.74 SD	120-127	7	90-96
7	+ .75 SD to +1.24 SD	112-119	12	78-89
6	+ .25 SD to + .75 SD	104-111	17	61-77
5	- .25 SD to + .25 SD	96-103	20	41-60
4	- .75 SD to - .25 SD	88- 95	17	24-40
3	-1.25 SD to - .74 SD	80- 87	12	12-23
2	-1.75 SD to -1.24 SD	72- 79	7	5-11
1	-1.75 SD and below	Below 72	4	4 & below

Fig. 2.7 Stanines for IQ's for a Group of Students with a Mean of 100 and an SD of 16

tribution, simply because the multiples of 4 are easier to remember and result in inconsequential differences in the converted scores. Note that only slight changes in the four percentages underlined are involved.

STANINE SCORES	1	2	3	4	5	6	7	8	9
PERCENT AT EACH LEVEL	4%	<u>8%</u>	12%	<u>16%</u>	20%	<u>16%</u>	12%	<u>8%</u>	4%

Stanines are often preferred in the interpretation of test data to students and parents because of the ease with which they are understood. Moreover, the coarseness of the unit reduces the chance of overgeneralization

on the basis of small differences in raw scores. Stanines also seem to be ideally suited for weighting and summarizing data of a wide variety of types, as in obtaining composite scores to be used in assigning marks, in homogeneous grouping, or in the selection of students for special classes.

As we have discussed normalized standard scores and the ways of obtaining them, the reader has probably become aware of a convergence among the types of scores discussed. That is, linear standard scores are almost identical with normalized standard scores when the distribution of raw scores approximates the normal curve (as it should when large numbers of students are tested with a test of appropriate difficulty). The reader has also seen that there is a definite relationship between percentile ranks and normalized standard scores. Furthermore, *T*-scores and *T*-scaled scores are seen as convenient transformations of their respective *z*-scores, used simply to eliminate the inconvenience of decimal points and negative numbers. Stanine scores are single-digit normalized standard scores, which are especially convenient for certain purposes.

The reader is now ready to study Table 2.7 and to examine Figure A.1 and to note not only the relationships already discussed, but a variety of other standard scores that have been developed for specific purposes.¹⁹

Converted Scores Based on the Average Grade or Age Status of Examinees Obtaining the Same Score

We will now consider the grade and age equivalents that are used so extensively with tests of achievement and scholastic aptitude. A converted grade or age score indicates the grade or age for which the student's test performance is typical. Age equivalents are stated in terms of the age (in years and months) for which test performance is typical, a mental age of 6-8 indicating a score on an intelligence test typical of a child 6 years and 8 months of age. A grade score or grade equivalent indicates the grade and months of attendance for which test performance is typical, a grade equivalent of 6.8 indicating a score typical of pupils who have attended sixth grade for 8 months. Here a decimal point is used, the year

¹⁹ The reader is warned, however, not to assume that CEEB (College Entrance Examination Board) scores are equivalent to the other scores shown, since the average score on students applying for college entrance would be considerably higher than that for persons in the general population; one could assume that the variability or *SD* would be less than for the general population. Also the AGCT score distribution would have proportionally fewer persons at the lower levels of mental ability than the general population. Hence, an AGCT score of 60 undoubtedly represents a higher IQ than 70. With these exceptions, however, the scores listed under each other in Figure A.1 are comparable.

Table 2.7
Major Types of Converted Scores for Tests Used by Educators and Psychologists^a

TYPE OF COMPARISON	MAJOR TYPES OF CONVERTED SCORES	TYPICAL NORM GROUPS	ADVANTAGES	DISADVANTAGES
1) Comparison with arbitrary standard	<p>1) Percentage correct</p> $X = \frac{\text{No. right} \times 100}{\text{No. of items}}$	None. Comparison with predetermined standard (as represented by perfect score on specific test)	<p>1) Not influenced by scores of other examinees</p> <p>Easily understood</p> <p>Provide basis for inferences concerning achievement on test content</p>	<p>Scores not comparable from test to test since tests vary in difficulty</p> <p>Provide no information on examinee's status with respect to others unless frequency distribution is examined</p>
2) Comparison with other examinees (in terms of difference between examinee's score and the mean—in SD units or some multiple thereof) ^b	<p>2a) z-scores (M equated to 0, SD to 1)</p> $z = \frac{\text{Raw score} - \text{Mean}}{\text{Standard Deviation}} = \frac{X - M}{SD}$	Single age, grade, or other group (such as 10th-grade girls or beginning file clerks), with which examinees can appropriately be compared	<p>2a) Make possible comparison of scores from test to test</p> <p>Differences in z-scores are proportional to differences in raw scores</p> <p>Can be used in computing averages, SD's, and correlations</p> <p>Frequency distributions of z-scores have same shape as distribution of raw scores on which they are based</p>	<p>Decimal points and minus signs are required</p> <p>Difficult to interpret without special training</p>

TYPE OF COMPARISON	MAJOR TYPES OF CONVERTED SCORES	TYPICAL NORM GROUPS	ADVANTAGES	DISADVANTAGES
	2b) <i>T</i> -scores (<i>M</i> equated to 50, <i>SD</i> to 10) $T = 10z + 50$	Same as above	2b) Advantages listed for <i>z</i> -scores plus the elimination of minus signs and decimal points	Meaning less obvious than for <i>z</i> -scores Size and range of values entail risk of confusion with <i>PR</i> 's
	2c) Deviation IQ's (<i>M</i> equated to 100, <i>SD</i> to 15 or 16, for example, $IQ = 15z + 100$ Stanford Binet $IQ = 16z + 100$)	Representative sampling of the general population at a specific age level	2c) Advantages listed for <i>z</i> -scores and <i>T</i> -scores Deviation IQ's have greater comparability from age to age than do ratio IQ's	Not directly comparable with standard scores obtained on achievement and other ability tests (that is, <i>SD</i> is 15 or 16, rather than 10, as for <i>T</i> -scores)
	2d) Standard scores on College Entrance Examination Board tests (<i>M</i> equated to 500, <i>SD</i> to 100) CEEB score = $100z + 500$	Students taking CEEB examinations in 1941	2d) Advantages listed for <i>z</i> -scores, plus elimination of minus signs and decimal points Small unit (0.01 <i>SD</i>) makes it possible for converted scores to reflect small differences in raw scores	Meaning less obvious than for <i>z</i> -scores Small unit increases hazard of user's attributing significance to unreliable differences between scores
3) Comparison with other examinees (in terms of rank of examinee's score within group)	3a) Percentile ranks (<i>PR</i> 's) percentile bands ^c	Single age, grade, or other group with which examinees can appropriately be compared	3a) Easily understood, even by persons untrained in measurement, such as students and parents Easily computed	Difficult to interpret as a measure of growth Cannot be used in computing averages, <i>SD</i> 's

Table 2.7 (Continued)
Major Types of Converted Scores for Tests Used by Educators and Psychologists^a

TYPE OF COMPARISON	MAJOR TYPES OF CONVERTED SCORES	TYPICAL NORM GROUPS	ADVANTAGES	DISADVANTAGES
	3b) <i>T</i> -scaled scores ^d (standard score equivalents of <i>PR</i> 's in a normal distribution) (<i>M</i> equated to 50, <i>SD</i> to 10) ^e	Same as above	3b) Similar to <i>T</i> -scores except that area transformations ^d make frequency distribution of <i>T</i> -scaled scores more nearly normal than distributions of raw scores on which they are based	Equal differences in <i>PR</i> 's do not represent equal differences in the raw scores from which they were obtained Risk of overemphasizing significance of differences between <i>PR</i> 's near median and underemphasizing significance of differences between <i>PR</i> 's near extremes Hazard of being confused with <i>PR</i> scores, which use same number range Not suitable when frequency distributions are definitely skewed, with a pile-up of scores at either end of score range

TYPE OF COMPARISON	MAJOR TYPES OF CONVERTED SCORES	TYPICAL NORM GROUPS	ADVANTAGES	DISADVANTAGES
3c) Stanine scores ^d (standard score equivalents of ranges in <i>PR</i> 's in a normal distribution) (<i>M</i> equated to 5, <i>SD</i> to 2) Each stanine unit except the two extreme ones is equal to 0.5 <i>SD</i>		Same as above	3c) Usable with any data that can be ranked Easily understood since interpretable in terms of ranges in <i>PR</i> 's All advantages of normalized standard scores, except that units are coarse Easily averaged and used in other computations Useful in interpreting test data to students and parents; risk of attaching too much significance to small differences in raw scores is minimized	Not widely used at present Coarse unit, having only nine different values
4) Comparison with other examinees (in terms of average age or grade status of examinees obtaining same raw score)	4a) Age scores ^e (mental ages, educational ages, and the like)	Successive age groups	4a) Apparently easy to interpret, especially for children in the preschool and elementary school years Useful in making inferences regarding child's readiness for certain learning experiences	Appropriate only for abilities showing continuous, relatively steady growth over several years Apparent obviousness of meaning can lead to generalizing beyond test and norm group

Table 2.7 (Continued)
Major Types of Converted Scores for Tests Used by Educators and Psychologists^a

TYPE OF COMPARISON	MAJOR TYPES OF CONVERTED SCORES	TYPICAL NORM GROUPS	ADVANTAGES	DISADVANTAGES
	4b) Grade-placement scores ^c	Successive grade groups (Grade placements may be based on full population, those of modal age, or those of modal age and intelligence)	4b) Uses a unit that is familiar to teachers Provides basis for inferences concerning the level at which groups are achieving on test content, as compared with representative norm samples	Since SD varies from test to test, scores that represent one year above or below CA in a series of subtests do not represent equal superiority (in terms of rank within norming sample) Disadvantages growing out of variations in SD indicated above Risk of misinterpretation in that student who shows a high degree of accuracy may be assumed to be able to work effectively two or three grades above his grade placement

Table 2.7 (Continued)

Major Types of Converted Scores for Tests Used by Educators and Psychologists^a

^a The classification used in columns 1 and 2 of this table is a simplified adaptation of that developed by Lyman, who includes many other converted scores in his comprehensive treatment of all such scores now in use. (Howard B. Lyman, *Test Scores and What They Mean*. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1963, pp. 90-135).

^b These derived scores are *linear* transformations of raw scores. A linear transformation is one obtained when the same constant is added to all scores and/or when all scores are multiplied by the same constant. The shape of the frequency distribution will be the same, whether one graphs the raw scores or their linear transformations. When a constant (C) is added to each score, the mean is increased by that amount (C), and the SD is unchanged. When each score is multiplied by a constant (C), the new mean and SD are C times as large as in the distribution of raw scores.

^c A frequency distribution for these derived scores (PR's, age, and grade scores) differs in shape from the frequency distribution of the raw scores from which they are obtained. In fact, a frequency distribution of percentile ranks is approximately rectangular, with almost equal percentages of scores in each interval. Moreover, equal differences in PR's (or age or grade scores) do not represent equal differences in raw scores. In the interpretation of PR's, one has to be especially cautious not to overemphasize the significance of differences

between PR's near the median (or midscore), and not to underemphasize the significance of differences between PR's near the extremes of the score range. PR's, age scores, and grade scores cannot be averaged or used in other computations that assume equality of units.

^d These derived scores are *area* transformations, since they are obtained by finding standard scores that correspond to cumulative percentages or areas in a normal distribution. That is, a raw score that exceeds 84 percent of the scores in the norming sample would be assigned a T-scaled score of 60, or a stanine of 7, because of the proportion of the area under the frequency polygon or curve that falls below that raw score. Unless the original distribution of raw scores is approximately normal, the shape of the distribution of normalized standard scores (such as T-scaled scores) will differ in shape from the frequency distribution of raw scores.

^e McCull, who developed the concept of the T-score, originally proposed that a value of 50 be assigned to the mean on a test for a group of unselected children twelve years of age; this specification, however, has been so frequently ignored that it has fallen into disuse. See William A. McCull, *Measurement* (New York: The Macmillan Company, 1939).

being divided into ten months on the assumption that growth in achievement takes place only during the ten months of the school year.

GRADE SCORES Although grade scores were not considered suitable for the history tests, they could be used for the arithmetic test since instruc-

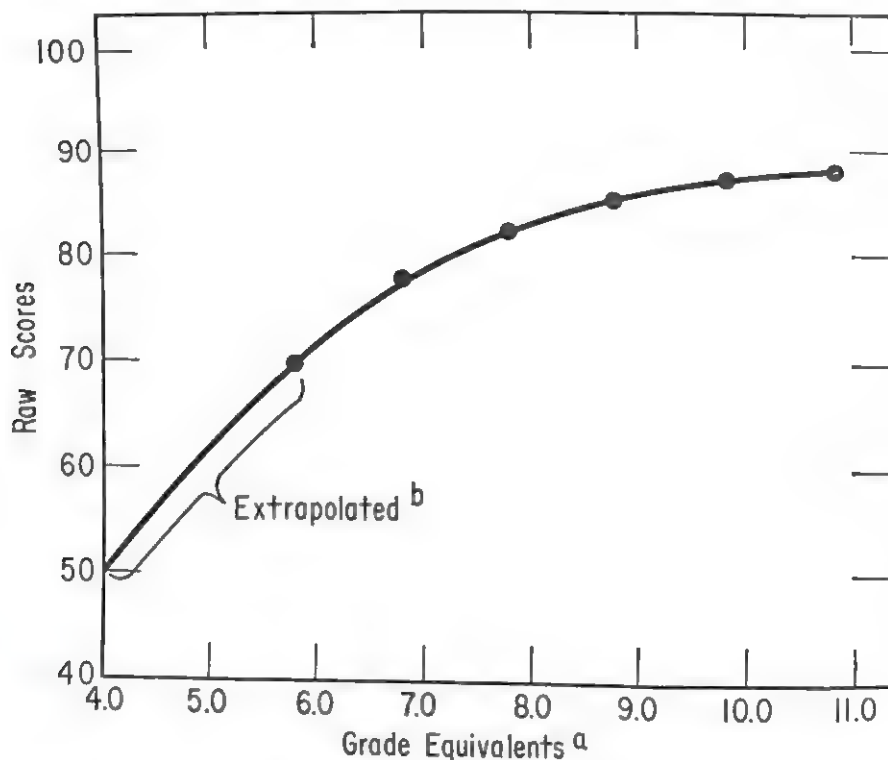


Fig. 2.8 Hypothetical Norm Curve for Obtaining Grade Scores or Grade Equivalents^a from Raw Scores on the Eighth-grade Arithmetic Test

^a Since the test was administered in the eighth month of the school year, the average raw-score value for the sixth grade is plotted above 6.8, the seventh grade above 7.8, and the like.

^b Usually a norm curve for a standardized test involves extrapolation, or extending the curve one or more years beyond the grades actually tested, using the general trend of the curve as a guide. Ordinarily, extrapolation should be limited to one grade below or one grade above the ones to which the test was actually administered. For example, in this test, the scores below 60 could be recorded as 5.0 —; the ability of students scoring so low would be more adequately measured by an easier test. It would be undesirable in this case to record grade scores above 9.0 because such a small difference in raw score is making a large difference in converted score. A more difficult test is needed to differentiate among superior students.

tion in arithmetic is given throughout the grades. To say that an eighth-grade student achieved as well in a comprehensive arithmetic test as the average pupil completing the seventh or eighth grade would be meaningful.

Developing Grade Norms for the Local Arithmetic Test. To develop local grade norms for this test, one would administer it during the same month of the school year to all students (or a random sample) in the sixth, seventh, eighth, ninth, and tenth grades. That is, it would be best to administer the test to representative students at the grade level for which the test is designed (in this case, eighth grade) and also to representative students one grade and two grades below, as well as one and two grades above. If the students were all tested at the end of the year, the graph relating grade score (or grade equivalent) to average raw score might resemble Figure 2.8. From a curve such as this, a table of grade norms could be constructed.

We would not be justified in obtaining grade norms for a test for a single grade level unless its content coverage were broad, including a considerable number of test items suitable for the younger and older students. Examination of an arithmetic subtest of a nationally used achievement test battery, such as the *Stanford Achievement Test* or the *California Achievement Test*, will illustrate such breadth of coverage.

The reader will recognize a problem involved in interpreting grade scores on this arithmetic test for students scoring at the ninth grade level and above. The meaning of a grade placement of 10.0 or 12.0 in arithmetic is difficult to interpret since no formal instruction in arithmetic is usually given beyond the eighth grade. Such grade equivalents in the mechanics of English would be meaningful because formal instruction in English continues throughout the high school years.

In the interpretation of age and grade norms, we tend to assume that we have scaled our raw scores to an external dimension (age or grade) that represents equal intervals. This assumption is often far from valid; for example, a year of growth in arithmetic from 2.0 to 3.0, or from 9.0 to 10.0, is far less than a year of growth from 6.0 to 7.0, because instruction in arithmetic is emphasized in the upper elementary grades. On the whole, grade norms are most suitable for use at the elementary school level, with tests of abilities that show a relatively steady growth over the years of instruction.

Grade Norms for a Published Achievement Test. A sample profile for the *California Achievement Test*, showing the data for a hypothetical student, is given in Figure 2.9. The grade equivalent for each of the subtests is given (following the student's score) and is also shown graphically on the profile. Since the student is in the eighth grade and was tested in March,

Fig. 2.9 Profile of California Achievement Test Results for a Hypothetical Junior High Student. Reproduced with the permission of the California Test Bureau from Ernest W. Tiegs and Willis W. Clark, California Achievement Tests, Junior High School Level (Monterey, Calif.: California Test Bureau, 1963).

his *actual grade placement*²⁰ (at the time of testing) is 8.6. That is, 8.6 is the national grade norm for him and his class.

It will be seen from the profile that this student is 7 months below his actual grade placement (or national grade norm) in total reading. He is 6 months below his actual grade placement in reading vocabulary and 8 months below grade norm in reading comprehension. He is 1.5 years above norm in arithmetic reasoning, and one year above in arithmetic fundamentals. His total language is one month below national grade norm.

Although the scores for minor subdivisions of each test (reading vocabulary in mathematics, science and the like) are plotted, one should not use these scores to read off grade equivalents at the right or left of the graph. Nor should one interpret small differences in these subscores as significant. As the reader will learn in the next chapter on reliability, scores on short tests are likely to fluctuate from one administration to another; and the interpretation of differences *between* scores is especially hazardous unless the scores compared are based on reasonably long tests of abilities that are not closely correlated with each other.

Problems in Interpreting Grade Scores. Unless one keeps in mind differences in the spread of achievement in different subject areas, misinterpretations can easily be made. Students at a certain grade level (for example, fifth grade) tend to be much more heterogeneous with respect to their achievement in reading and language usage than in such subjects as arithmetic, where the student's progress is more arbitrarily controlled by curricula and textbooks.

It is evident from Table 2.8 that a class which is generally superior, in the sense that its average score in each subtest exceeds 75 percent of students in the norming population, would appear to be doing much better in reading and language usage than in arithmetic. For a student or

²⁰ The actual grade placement for a child or a class group at time of testing is determined by the *grade* in which the pupils are enrolled and the *date* on which the test is given. A table for determining actual grade placements is usually given in the test manual. The following is quoted from the manual of the *California Achievement Test*:

Actual grade assignment is determined by adding to the pupil's grade the following fractions of a year:

<i>Months</i>	<i>Low Section</i>	<i>High Section</i>
September or February	.0	.5
October or March	.1	.6
November or April	.2	.7
December or May	.3	.8
January or June	.4	.9

Where schools have annual promotions only, ignore the Low Section and High Section captions.

class that is generally inferior, in the sense of exceeding only 25 percent of the norming population on each subtest the reverse would seem to be true, that is, higher achievement in arithmetic than in reading and language usage. Both situations are attributable to the greater homogeneity of students in arithmetic achievement. Similar comparisons for two other widely used achievement tests (the *California Achievement Test* and the *SRA Achievement Series*) showed similar results, with much greater heterogeneity in reading and language usage than in arithmetic.²¹

Table 2.8

Grade Placements (GP's) for a Beginning Sixth-Grade Student or Class Achieving at a Generally Superior or Generally Inferior Level on the *Stanford Achievement Test*

Subtest	GP's for student or class achieving at	
	P_{75} LEVEL	P_{25} LEVEL
Reading	7.7	5.1
Language	8.0	4.9
Arithmetic	7.0	5.4

Source: Warren G. Findley, "Use and Interpretation of Achievement Tests in Relation to Validity," *18th Yearbook, National Council on Measurement in Education* (Ames, Iowa: The Council, 1961), p. 32.

This comparison indicates that scores which are equal in terms of one norms table may not be equal in terms of another. Hence, one must be cautious about reading into converted scores more than is really implied. When a sixth-grade student scores 8.0 in language and 7.0 in arithmetic on the *Stanford Achievement Test* for grades 5-6, we are justified in saying that he did as well on this language subtest as did the average eighth-grader in the norming sample and as well in the arithmetic subtest as the average seventh-grader in that group. We cannot infer, of course, that he would do as well in those subjects as the grade scores imply if he were given a test designed for the junior high school level. He might, but we are certainly not justified in going that far beyond the data. Moreover, his achievement in language and arithmetic is *equal*, or at the same level, if the comparison is made in terms of percentile scores, or the percentage

²¹ Warren G. Findley, "Use and Interpretation of Achievement Tests in Relation to Validity," *Eighteenth Yearbook, National Council on Measurement in Education* (Ames, Iowa: The Council, 1961), pp. 32-34.

of sixth-graders that he exceeds (Table 2.8), or if comparisons are made in terms of normalized standard scores, which are derived from percentile scores.

The hazards involved in interpreting grade and age norms are sufficiently great that the following recommendations are made in the "Technical Recommendations for Achievement Tests."

F 2 Where there is no compelling advantage to be obtained by reporting scores in some other form, the manual should report scores for defined groups in terms of percentile equivalents or normalized scores for defined groups.
VERY DESIRABLE

F 2.1 If grade norms are provided, tables for converting scores to percentiles (or standard scores) within each grade should also be provided. ESSENTIAL²²

Hence, it is evident that even though teachers feel at home with grade and age norms, it is best to supplement these norms by some type of converted scores that reflect the student's rank *within* his grade or some other appropriate reference group. Such supplementation is especially important for students or groups that deviate considerably from the average of their age or grade level in the abilities tested.

For some kinds of decisions (such as deciding on the reading level of instructional materials for a student, or deciding whether he has made normal progress during a school year), we are interested in knowing what group the student most resembles—where he stands on the "ladder" of mental growth or educational achievement. In these situations, it is meaningful to use grade or age norms. In other cases, for example, when we want to study a student's relative strengths for vocational guidance, we need to compare the student with those of *his own* age or grade level, college major, or vocation; in these cases, percentile ranks or standard scores seem to serve the purpose better. For other decisions, such as grouping students *within* a school or selecting students for a scholarship, any type of converted scores, or even raw scores, are adequate for placing students in rank order. Whenever the ranking of students should be based on a composite of several scores, so that decisions can be made on the basis of as much information as possible, the use of some type of standard score (especially stanine scores) is advisable.

AGE SCORES Age norms are obviously useful during the preschool years for various evidences of developmental maturity; for example, to say

²² Committee on Test Standards, American Educational Research Association and National Council for Measurement in Education, *Technical Recommendations for Achievement Tests* (Washington, D. C.; National Educational Association, 1955), p. 34.

that the child has as good balance or finger dexterity as a four-year-old is quite meaningful. In fact, age norms for development in motor skills, or any other aptitude highly correlated with age, are valuable.

The age equivalents of raw scores on an achievement battery are called educational ages, while those for a reading test would be called reading ages. Such age scores are infrequently used in interpreting achievement test data. In fact, the chief use of age scores has been on tests of scholastic aptitude.

Originally developed in connection with individual tests of mental ability, mental ages have served as a meaningful type of norm, especially as data have cumulated from research studies concerning the mental ages apparently required for the achievement of various types of educational and vocational tasks. As we shall see in Chapter 6, however, the assumption of equality in units of mental age holds much better during the elementary school years than it does during adolescence; and the concept of MA becomes meaningless as an attempt to interpret the level of scholastic aptitude of the superior adolescent, or that of the normal or superior adult.

Special Types of Converted Scores Designed to Serve Special Purposes

LONG-RANGE EQUAL-UNIT SCALES DESIGNED TO MEASURE GROWTH Although age and grade equivalents have serious limitations, they do provide a basis for making inferences concerning how far students and groups have proceeded up an educational or developmental "ladder." Standard scores, despite their points of superiority, are always based on comparisons *within* a defined age, grade, or other group.

A few attempts have been made to develop scores that have some of the advantages of standard scores but are designed to allow the teacher to make inferences about student progress up the "ladder." The *K*-score, used with the 1953 edition of the *Stanford Achievement Test*, is one such attempt. The *K*-scores for this test are so designed that the average performance of tenth-grade students is equated to a *K*-score of 100; the unit of measurement (that is, a score unit of 1) is equated to $\frac{1}{4}$ of the *SD* of the scores of fifth-grade pupils. Differences between pairs of *K*-scores are therefore comparable throughout the range of grades measured by this test battery. Further research on systems of this kind would seem highly advisable since they combine the advantages of the "position-on-the-ladder" approach with the comparability and meaningfulness of standard scores.

STANDARD SCORES DESIGNED TO BE COMPARABLE FROM ONE HIGH SCHOOL SUBJECT TO ANOTHER Although percentile scores and normalized standard scores seem most feasible for use at the high school level,

they can also be misleading unless some adjustment is made for the differences in scholastic aptitude among students who *elect* the various high school subjects. If a student scores at the 80th percentile in a biology test and the 40th percentile in a physics test, one would seem justified in concluding that the student showed a marked superiority in biology as compared to physics. However, almost all students take required courses in biology, while only above-average students elect physics and only the superior students survive to take an end-of-course examination. Hence, it becomes impossible to make an inference about this student's relative competency in these two fields unless the norms are based on comparable groups of students.

At least two publishers have developed systems of scaled scores for a series of high school tests, designed to cope with this problem. In a system of converted scores used by the Cooperative Test Division, Educational Testing Service, 50 is defined as the score that would be made by a student of *average ability* who had had typical instruction in the course. In another system (utilized in the *Evaluation and Adjustment Series* of tests in high school subjects, published by Harcourt, Brace & World), the *M* and *SD* for the standard scores on each subject-matter test are set at the same level as the *M* and *SD* for intelligence quotients for the representative group of subject enrollees on which the test was standardized. For example, if students in the norming sample taking a specific mathematics test had an average IQ of 109 and an *SD* of 12 IQ points, the standard scores for that mathematics test would be established with an average of 109 and an *SD* of 12. This system also facilitates comparison between a student's achievement and his capacity; for example, a student who made an average score of 109 in this mathematics test but had an IQ of 125 could readily be identified as working below ability level.

DEFINING NORMING POPULATIONS AND SELECTING NORM SAMPLES

Norms for the Population-in-General vs Norms for Homogeneous Groups

In interpreting test data on average school achievement in the fundamental skills or in different content areas, general-population norms are needed as a basis for comparison. In counseling a student about educational and vocational plans, however, general-population norms may be of limited value. Norms are needed that are relevant to the inference one wishes to make, for example, norms for "freshmen accepted by engineering schools" for a student aspiring to that vocation.

Whether general-population norms or norms for selected homogeneous

groups are provided, it is important that the test authors define the populations to be sampled and then use effective procedures to see that the norming samples represent those populations.

A test author, for example, might develop an entrance examination for private liberal arts colleges. In such a case, his norming sample need not be representative of students in general, or even of college freshmen. If he defined his population as "applicants for freshman standing in liberal arts colleges," he could proceed to sample that defined population.

In standardizing vocational aptitude tests, several populations may be sampled to increase the number of valid inferences that can be made from test scores. A test of clerical aptitude, for example, should be normed on several relatively homogeneous groups of people with whom it is sensible to compare the student's performance. For example, a girl taking a secretarial course in a commercial high school would like to know how her score of 152 on the *General Clerical Test* compares with the scores of those with whom she will be competing for employment. Examination of the norms for this test reveals that her score of 152 can be compared with two relevant groups, as follows:

PERCENTILE RANK	NORMING SAMPLE	INTERPRETATION
PR 85	Commercial high school senior girls	Her score exceeds that of 85 percent of commercial high school senior girls in general
PR 75	Commercial high school senior girls completing secretarial training	Her score exceeds 75 percent of commercial high school senior girls completing secretarial training

In addition, the local school district has obtained information on the test scores of applicants to two large local firms that employ their graduates. In comparison with firm X's clerical applicants, Sue's PR is 60, while in comparison with their secretarial applicants, her PR is only 32. The other large firm in town, firm Y, has not compiled its own norms, but has merely established critical scores²³ on this and other tests, below which applicants will not be considered. For clerical applicants, firm Y has set a critical score of 130, for stenographic applicants, its critical or cut-off score is 150.²⁴

²³ A critical score or a cut-off score is a score below which applicants are rejected as being too low on some critical qualification. For example, the police force will have critical scores with respect to height and vision; people scoring below these critical or cut-off scores are not considered further in the selection process.

²⁴ These case data are excerpted from "Norms Must Be Relevant," *Test Service Bulletin No. 39* (New York: The Psychological Corporation, 1950).

From all these data, it is apparent that although this student is quite superior to her fellow-students (exceeding 85 percent), she excels only three-fourths of the secretarial graduates. According to the norms and cut-off scores of local firms, she would be considered above-average for clerical applicants but marginal as an applicant for stenographic work, insofar as the score on this test is concerned. Obviously factors other than clerical proficiency are also considered in the employment and retaining of applicants; and her score is sufficiently high that these other factors will probably be taken into account.

Figure 2.10 indicates clearly the importance of using such special norms to aid in interpreting tests used in vocational guidance or tests administered as aids in the selection and assignment of employees. Although scores on the number-checking section of the *Minnesota Clerical Test* are normally distributed for workers in general, with relatively few obtaining A

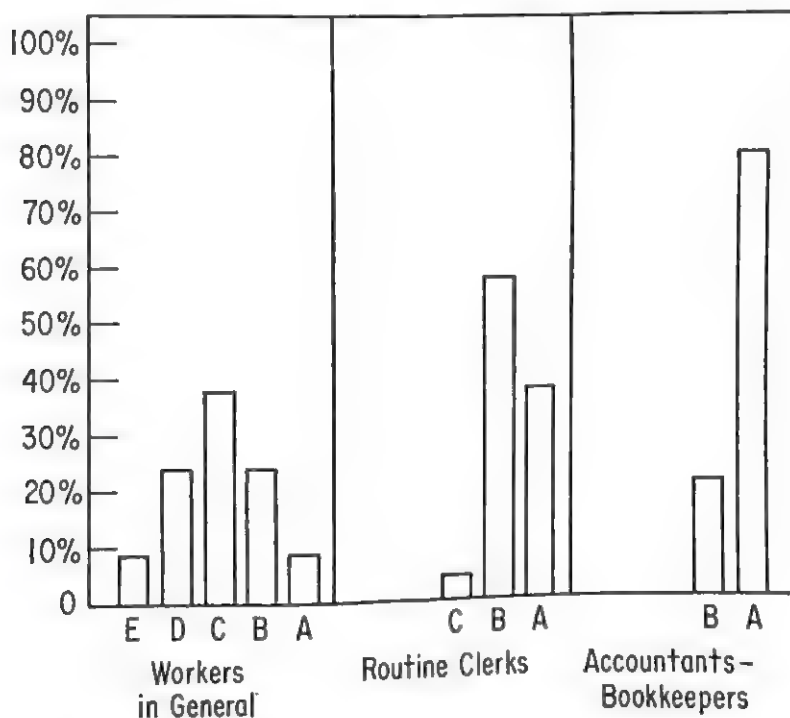


Fig. 2.10 Occupational Differences on the Minnesota Clerical (Numbers) Test Showing the Percentage of Each Type of Worker Making a Given Letter Grade.

From D. M. Andrew and D. G. Paterson, "Measured Characteristics of Clerical Workers," *University of Minnesota Bulletin of the Employment Stabilization Research Institute*, Vol. 1, 1934.

and B scores, almost all clerks have A and B scores and almost all accountant-bookkeepers have A scores. In fact, an 18-year-old high school senior whose raw score of 106 would give him a *PR* of 58 for his age group and a *PR* of 74 for "all employed adults" would obtain a *PR* of only 1 for accountant-bookkeepers.²⁵

Procedures Used in Obtaining Norming Samples Representative of Norming Populations

One of the questions frequently asked in measurement classes concerns the number of cases needed for an adequate norms table. Actually one can test many thousands of cases that happen to be available and still not have a representative sample of the defined norming population. Bias (usually in the direction of superior achievement) is often involved when the publisher uses those test data that test users send in voluntarily or when he tests only those intact classes that school principals suggest as good classes to test.

With respect to number of cases, the sample should be sufficiently large that the norm curve drawn would not differ significantly from one that would be drawn on the basis of another sample of the same size. Prior to our discussion of reliability in the next chapter, all we can say is that the size of norming samples and subsamples should be sufficient to provide stable values; that is, the converted scores for specific raw scores should not fluctuate significantly if the test were renormed on another sample of the same size.

When a published test of achievement or aptitude is standardized to obtain national norms, the tests must be administered to norming samples that are representative of students at each age and grade level (for which the test is designed) in the country as a whole. The problem of obtaining norming samples that are representative of age or grade groups within the population-in-general has always been a very difficult one. School children of a specific age level, for example, age 11, are in different grades. The difficulties of locating representative age groups of preschool-age children are even greater.

The author of a high school test will have great difficulty in obtaining a representative sampling of 17-year-olds. He must either make extraordinary efforts to locate drop-outs, as well as accelerated students already in college, in order to obtain a representative sampling of all 17-year-olds; or he can choose to define his norming population as "students attending high school," and then proceed to sample that population.

The number and representativeness of the *communities* included in the

²⁵ Donald E. Super, and John O. Crites, *Appraising Vocational Fitness* (New York: Harper & Row, Publishers, Inc., 1962), p. 166.

norming population tested are especially important in norming an achievement test; a very large number of cases from only a few communities may result in distorted standards because of differences in curricular emphasis and achievement in various parts of the country.

In the standardization of the *Stanford Achievement Test*, every student from certain grade levels in 340 communities was tested; then the norms were developed on the basis of a much smaller random sampling of these students. The 340 communities included 104 from the New England and Middle Atlantic states, 59 from the North Central states, 58 from the Southern states, and 119 from the Pacific Coast states. One hundred of the communities included were small towns under 2500 population; one hundred ranged from 2500 to 24,999; 104 were county, district, or union school districts; and 36 were communities of 25,000 or more. This norming procedure is an example of superior practice in obtaining representative samples.

A growing body of information has been developed on characteristics of school systems that are related to test performance. The United States Bureau of the Census maintains a card file of approximately 70,000 school systems, with sufficient information concerning them, that one can draw a sample of school systems that is reasonably representative with respect to factors related to school achievement or scholastic aptitude. The following statement concerning the standardization of the *Henmon-Nelson Tests of Mental Ability* illustrates good procedures:

A multi-stage sample of school systems was obtained, yielding 250 systems stratified by size of system (with 8 size classifications) and by geographical region. School directories for the 48 states were consulted so that an elementary school within each of the 250 systems could be drawn with probability proportionate to school size.²⁶

The work of obtaining national norms will be greatly simplified when we have available the results for "anchor tests" administered to truly representative samples of the population-in-general. Through Project Talent and the work of the American Textbook Publishers Institute, anchor tests are being developed, which can be administered along with new tests to norming samples and thus assist in the development of converted scores that are more nearly comparable from test to test.²⁷ Project Talent has involved

²⁶ Tom A. Lamke, "The Standardization of the Henmon-Nelson Revision," *The Thirteenth Yearbook of the National Council on Measurements Used in Education* (Ames, Iowa: The Council, 1956), p. 43.

²⁷ Lee J. Cronbach, *Essentials of Psychological Testing*, 2d ed. (New York: Harper & Row, Publishers, Inc., 1960), p. 93; Roger T. Lennon, "Discussion of the School Administrator's Problems," *Invitational Conference on Testing Problems* (Princeton, N. J.: Educational Testing Service, 1957), p. 98.

the testing of a random sampling of 5 percent of all students in grades nine through twelve. Calibrating new tests against these anchor tests may achieve new levels of comparability in converted test scores.

The publishers of the *Stanford Achievement Test* and the *Metropolitan Achievement Test* have introduced a significant refinement into the development of grade norms. They have based their "modal-age grade norms" on the scores of all students in a given grade who are typical with respect to age. By means of these norms, the achievement of a student can be compared with that of students who make normal progress in school. The modal-age group tends to be slightly higher in ability than an unselected group, in that larger numbers of dull overage than bright underage students have been excluded. Hence, the publishers contend that modal-age grade norms provide a better standard of accomplishment for students than norms based on the total grade population.

National Norms as Standards

One of the disconcerting phenomena in recent years is the extent to which national norms have been used as standards of accomplishment. The national norms for a test are a set of organized data that show the test scores earned by norming samples representative of defined populations. These norms should aid the user in making sensible inferences regarding (1) the present performance of individuals and groups on each test and (2) their relative level of performance in different subjects or characteristics.

Sometimes a teacher uses the grade norm as the standard each student *should* be able to achieve, with little allowance made for how each student compares with the norming sample with respect to scholastic aptitude or his mastery of prerequisite skills or information. Many test manuals provide information concerning scores earned by subgroups in the norming population who are above average or below average in scholastic aptitude. Norms for such groups are of considerable assistance in the interpretation of student achievement of individuals and groups that are atypical. In fact, the *California Achievement Tests* provide *anticipated achievement grade-placement* scores, so that the achievement of each pupil in each test can be compared with pupils in the norming sample with the same age, grade placement, and mental age. These AAGP scores can be thought of as expectancy scores, which indicate the average achievement of children comparable to the examinee with respect to chronological age (CA), mental age (MA) and actual grade placement.

A quite different problem arises from the fact that many schools have become complacent because their students are up-to-norm on all skills. There should be little reason for satisfaction about average achievement by

classes with superior scholastic aptitude. For that matter, there is little basis for gratification when classes with average intelligence achieve at national norm when one considers the many factors (in the student and the learning situation) that tend to keep students in the nation as a whole from achieving at an optimum level.

Dressel has urged that norms be based on student achievement under optimal conditions so that schools could set for themselves standards that are not based on mediocre accomplishment. Such norms, of course, would have to be supplementary to the types that are now available.²⁸ These norms would be especially appropriate, as Dressel implies, in areas in which schools give lip-service to new objectives but do not proceed to implement the objectives effectively.

USE OF DIFFERENT TYPES OF CONVERTED SCORES IN COMPUTATION

Throughout this chapter, various statements have been made concerning computation with converted scores; for example, the reader was warned that he should not average percentile ranks. It might be well at this point to consider the four different types of number systems.

1. *Nominal* numbers are used as symbols for categories. Illustrative of *nominal* numbers are area telephone codes or code numbers assigned to designate different areas of interest in an interest inventory. For such numbers, the term "greater than" or "less than" have no meaning. These numbers are countable, but cannot be ordered or ranked. However, frequency distributions by categories can be prepared and analyzed by suitable statistical techniques.
2. *Ordinal* scales involve numbers that can be counted and ranked but should not be used in computing sums, differences, or ratios; in an ordinal number system, differences of equal size do not have comparable meaning at different points in the scale. If there are 20 students in a class, each student can be assigned a rank with respect to any defined characteristic; or students' essays or art products can be ranked from 1 to 20 with respect to their general quality or some aspect thereof. However, one cannot infer that the differences in quality between ranks 1 and 2 is the same as the difference between ranks 11 and 12. An ordinal scale reflects position in an ordered series, but the scale does not have equality of units.

Ranks and *PR*'s are ordinal numbers. Because of the lack of comparability of units at different points in the scale, age and grade equivalents should also be considered as ordinal numbers. For these types of converted scores, we should use only formulas that make no assumptions about equality of units.

²⁸ Paul L. Dressel, "The Way of Judgment," *The Fifteenth Yearbook of the National Council on Measurements Used in Education* (New York: The Council, 1958), pp. 5-8.

For example, one can legitimately compute the median rank or *PR*, but one should not compute a mean, for its computation assumes equality of units. The quartile deviation²⁹ should be used as a measure of variability, rather than the standard deviation, which assumes equality of units. Measures of relationship must be obtained by special means suitable for ordinal numbers (see Chapter 3).

3. *Interval scales* are based on equal or comparable measurement units throughout the scale; that is, an interval of 1, 2, or more points has the same meaning throughout the scale. Interval scores cannot only be counted and ranked but used to compute sums of scores and differences between scores. One can make inferences on the basis of these sums and differences. Interval scales, however, have no meaningful zero point. For example, a pupil making zero on a very difficult spelling test does not have zero spelling ability; he might make a fairly high score on an easy test. In fact, we devise many tests in such a way that the easy items are omitted to reduce test length; we do not test junior high school students on simple spelling words that almost all students have learned during their early school years.

Nunnally illustrates an interval scale by the example of obtaining data on relative running times in a race (as one second *behind the winner*, two seconds, three seconds, and the like).³⁰ We could average these data and perform any kind of computations with them that were based on differences between scores. However, we could not say, on the basis of these scores, that one runner was twice as fast as another. Our thermometers have an interval scale with equal units but no meaningful zero point. In fact 0° on the Centigrade scale represents 32° on the Fahrenheit scale; moreover 0° on neither scale represents the lowest possible temperature. Hence we cannot say that 60° on either scale is twice as warm as 30°.

If we consider raw scores on tests to represent "number of questions answered correctly,"³¹ we can treat these scores as interval numbers; these scores can therefore be used in the computation of means, *SD*'s, and measures of relationship, to be explained in Chapter 4. However, we cannot use these scores to compute ratios; that is we cannot say that a test score of 60 is twice as good as one of 30.

4. *Ratio scales* have not only equal measurement units but also a meaningful zero point. Ratio numbers are countable and rankable; they cannot only be

²⁹ The quartile deviation (*Q*) is equal to one-half the range between Q_3 (or P_{75}) and Q_1 (or P_{25}). The formula is: $Q = \frac{Q_3 - Q_1}{2}$.

³⁰ J. C. Nunnally, Jr., *Tests and Measurements: Assessment and Prediction* (New York: McGraw-Hill Book Company, Inc., 1959), p. 11.

³¹ When we consider test scores as representing an attribute of an individual, it is apparent that in this sense the scores do not represent an interval scale but only an ordinal scale (since equal differences in raw scores do *not* represent equal differences in the attribute). For example, a 5-point difference between spelling scores of 50 and 55 on a 100-item spelling test may not represent nearly as large a difference in spelling ability, as a 5-point difference between scores of 95 and 100. The students with perfect or near-perfect scores may not have been able to show their superiority on this test. However, raw scores and their linear transformations are ordinarily considered in the sense of "number right" and hence are interval scores for purposes of computation.

used in computations that assume equal differences between successive numbers; but *ratios* between such numbers can be computed. That is, when we measure running time in seconds, we can say that one runner is twice as fast as another. Most scales in physical measurement are ratio scales. Hence we justifiably use them to compute ratios; we can say that one man is 1.25 times as tall as another; that one man has walked twice as far as another.

In testing, we obtain ratio scores only when we test a random sampling of a precisely defined universe of items. For example, when we develop a test that is a random sampling of *all* spelling words studied, we can say that Tom, who has a score twice as high as Sue, can spell twice as many of *this universe* of spelling words. We could make similar inferences about tests involving random samplings of the addition, subtraction, multiplication, or division facts.

Of course, the errors of measurement³² (which reflects variations in sampling of words and temporal fluctuations in student performance) must be taken into account in judging the confidence with which such a statement can be made. However, this type of score represents the closest approach (in measurement) to ratio scores. Note that in interpreting student performance on a sample of spelling words, we do not generalize beyond the population of words sampled, in this case the words in the state speller for that grade level. In such a situation, a meaningful zero point exists.

In some measurement textbooks, the statement is made that raw scores are meaningless; one can see that in a situation such as the spelling test described, raw scores can serve as the basis of meaningful inferences. Moreover, in many situations, where the user is concerned only with the student's rank within a group (such as in assigning marks or in selecting students for some award), raw scores are just as usable as converted scores.

SUMMARY STATEMENT

Before a student's scores on different tests can be interpreted as indicating relative strengths and weaknesses, they must be translated into converted scores, which indicate how his test performance compares with those of others in some reference group with which he can be appropriately compared.

Three major approaches to obtaining converted scores, on the basis of comparisons with appropriate reference groups, were studied:

1. Standard scores (*z*-scores, *T*-scores, and the like, based on the difference of a student's score from the group average, expressed in *SD* units or some multiple thereof)
2. Percentile scores, normalized standard scores, or stanine scores (based on the relative position, or rank, of the student's score within the group of all students tested, or some defined reference group)

³² Measurement errors are discussed in the next chapter.

3. Age or grade scores (the average age or grade status of students obtaining the same score)

Each of these approaches was illustrated with local data compiled to answer questions with respect to student achievement in spelling, arithmetic, and history.

The advantages and disadvantages of each type of norms were summarized in Table 2.7. In this table, the formula or the procedures for computation are given for each type of norm, as well as the typical reference groups used as a basis of comparison.

SELECTED REFERENCES

- EBEL, ROBERT L., "Content Standard Test Scores," *Educational and Psychological Measurement*, vol. 22 (Spring 1962), pp. 15-25.
- ENGELHART, MAX D., "Obtaining Comparable Scores on Two or More Tests," *Educational and Psychological Measurement*, vol. 19 (Spring 1959), pp. 55-64.
- FRANZBLAU, A. M., *A Primer of Statistics for Non-Statisticians*. New York: Harcourt, Brace & World, Inc., 1958.
- GARDNER, ERIC F., "Value of Norms Based on a New Type of Scale Unit," *Proceedings, 1948 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 1949, pp. 67-74.
- LYMAN, HOWARD B., *Test Scores and What They Mean*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1963.
- NEDELSKY, LEO, "Absolute Grading Standards for Objective Tests," *Educational and Psychological Measurement*, vol. 14 (Spring 1954), pp. 3-19.
- SEASHORE, HAROLD G., "Methods of Expressing Test Scores," *Test Service Bulletin*, No. 48. New York: The Psychological Corporation, 1955. Available on request.
- _____, AND JAMES H. RICKS, JR., "Norms Must Be Relevant," *Test Service Bulletin*, No. 39. New York: The Psychological Corporation, 1950. Available on request.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. How should national norms on achievement tests be used? Evaluate the following uses of national grade placement norms:
 - a. As a standard by which a school's average achievement is judged.
 - b. As a standard to be attained by each individual.
 - c. As a norm which should be surpassed by the upper half of the class.
 - d. As a norm which should be applied to each individual in relation to some measure of his learning ability.
 - e. As a norm which represents the average accomplishment of pupils in each grade who are of average intelligence.
2. List three typical situations in which achievement test norms can be used to advantage in interpreting test results.

3. Why are age and grade scores unsuitable for tests in most high school subjects? What are the relative advantages of standard scores and percentile scores?

4. Using the mean and *SD* given in Table 2.3, convert the following raw scores into *z*-scores:

$$X = 95, X = 75, X = 65.$$

5. Using Table 2.6, find the percentile ranks, *T*-scores, and stanine scores for the following raw scores:

$$X = 55, X = 65, X = 75, X = 85.$$

6. For each of the following sets of test scores for the sixth-graders of a school system, select a suitable size of interval, and set up a form for tallying the scores:

TEST	RANGE OF SCORES
Spelling	9- 49
Vocabulary	18- 98
Interest inventory	70-240

7. Tally the reading grade placements in Table 14.1. Compute P_{50} (the median score), P_{75} and P_{25} .

8. Would national norms be necessary for interpreting:

- A test of ability to understand oral Spanish which is given as the chief basis of admitting students to a class in Spanish conversation?
- An intelligence test given to children being placed by an adoption service?
- A diagnostic reading test given to determine whether students need remedial instruction?
- An aptitude test being used in the vocational guidance of high school students?

When a person takes a test, we obtain a limited sampling of his performance in the area tested. Two different forms of a test, even though they are designed to be equivalent, provide somewhat different samples of behavior. Moreover, individuals vary from one test session to another in their level of motivation, speed of working, and other characteristics.

The concept of *reliability* has to do with consistency of measurement, the extent to which an individual's scores vary from one sampling to another of the same type of behavior. When we are making decisions concerning individuals, we need to obtain many samples of behavior. As Diederich says,

There is very little hope of proving anything in education with single measures. The real hope lies in repeated measurements: either testing many students with each single measure, or testing the same student with many different measures. Hence, we like to have repeated measurements of scholastic aptitude, reading achievement, and other important variables if we are to draw inferences concerning individuals.¹

Variations in an individual's test behavior on different samplings of items and different testing occasions sometimes result in overestimates, and sometimes in underestimates, of the examinee's ability. These errors are *compensating errors*; they tend to average out in repeated testings. Other types of errors (such as a student's habit of cheating on tests, or his characteristic carelessness in reading test directions) constitute *systematic errors*, which

¹ Paul B. Diederich, "Short-cut Statistics for Teacher-Made Tests," *Evaluation and Advisory Service Series No. 5*. (Princeton, N. J.: Educational Testing Service 1960), p. 20.

do not average out but tend to systematically raise or lower an individual's scores. In this chapter, we are concerned chiefly with the types of compensating errors involved in obtaining limited samplings of behavior. Systematic errors are discussed more fully in Chapter 4 on validity.

INTERPRETING TEST SCORES IN TERMS OF SOURCES OF VARIANCE

Individual differences with respect to test scores arise from many sources (Table 3.1). Our aim in test construction and administration is to have most of the variation in test scores attributable to individual differences in the ability or trait we wish to measure. We recognize, however, that much of the variation in test scores is due to other factors, for example, individual differences in speed of working, "testwiseness," and other factors that affect a person's scores on many tests.

Many of the converted scores discussed in the preceding chapter are based on a measure of dispersion or variability of scores among individuals, that is, the standard deviation. Although an approximation formula for the *SD* was used in Chapter 2, we should now learn the basic formula.²

$$SD = \sqrt{\frac{\sum x^2}{N}} \text{ where } x = X - M$$

The standard deviation can be computed by finding all *x* values (or deviations from the mean) and substituting them in the formula. Or one can use the short method with grouped data, as shown in Table A-3 in the Appendix.

The expression under the radical sign $\frac{\sum x^2}{N}$ is called the "variance" (*V*). The term "variance" is an expression for the amount of scatter around the

² If we want to estimate the variability of the *population* that a sample represents, the formula should read, $\sigma = \sqrt{\frac{\sum x^2}{N-1}}$. Conventional practice calls for the use of letters, such as *M*, *SD* (or *s*), for values obtained for specific samples; while corresponding Greek letters (μ , σ) are used for estimates of the mean and *SD* for the population, from which a random sample has been taken. Most textbooks use Greek letters in the presentation of formulas because it is assumed that the investigator is interested, not in the *SD* for his sample, but in σ (an estimate of the standard deviation for the population which the sample represents). In many studies in education, however, the researcher is not interested in generalizing beyond the sample tested to the population that the sample represents. Hence, we have avoided the use of Greek letters in this textbook; with few exceptions, we are discussing data obtained from actual samples.

Table 3.1
Possible Sources of Variance in a Test Score

Variance due to the <i>LS</i> and <i>TS</i> categories is measurable by comparisons of examinees' scores on different forms or test samples	<i>LG</i> —Relatively <i>lasting</i> , ^a <i>general</i> ^b characteristics of the examinee	Variance due to categories <i>TS</i> and <i>TG</i> is measurable by comparison of examinees' scores on different testing occasions
	<i>LS</i> —Relatively <i>lasting</i> characteristics of the examinee elicited by this <i>specific</i> test sample	
	<i>TS</i> — <i>Temporary</i> characteristics of the examinee elicited by this <i>specific</i> test sample	
	<i>TG</i> — <i>Temporary</i> but <i>general</i> characteristics of examinee (likely to affect his performance of any tests given on that occasion, for example, a series of tests given on a single occasion for civil service testing or admission to college)	

Illustrative Sources of *LG*, *LS*, *TS* and *TG* Variance

***LG*—(LASTING GENERAL) VARIANCE**

1. Ability to respond successfully to stimuli of the type presented in this test
2. Ability to comprehend and follow directions, "testwiseness"
3. General abilities useful in many tests (for example, reading, perceptual speed, memory)
4. Attitudes, habits, or emotional reactions that characterize the examinee's behavior in situations like the test situation (for example, self-confidence, anxiety, tendency to guess when uncertain, tendency to give socially approved answers)

***LS*—(LASTING SPECIFIC) VARIANCE**

1. Knowledges and skills required in this specific test sample (for example, knowledge of how to spell specific words, accuracy with specific number combinations)
2. Characteristic examinee attitudes, habits, or emotional reactions elicited by this specific test sample (for example, a general tendency to feel tense and anxious, "triggered" early in the test by the inclusion of items not covered in local course of study)

***TS*—(TEMPORARY SPECIFIC) VARIANCE**

1. Fluctuations in memory for particular facts
2. Level of practice, or recency of review, on skills and knowledges tapped by this test sample (for example, differential effects of special coaching on examinee's success on a specific set of items as compared with another set)
3. Variations in examinee's attention, degree of concentration, speed of response, or standards of judgment (resulting from factors specific to the test sample, such as the examinee's interest in the particular reading selections or science problems included in the sample)

Table 3.1 (Continued)
Possible Sources of Variance in a Test Score

-
4. Temporary emotional states related to particular test stimuli (for example, a question that calls to mind an upsetting disagreement with someone in authority)
 5. Luck in the selection of answers by "guessing"

TG—(TEMPORARY GENERAL) VARIANCE

1. The examinee's condition on that testing occasion, with respect to such factors as health and emotional strain
 2. Effects of such conditions in the testing environment as heat, light, ventilation
 3. Level of motivation, as affected by his perception of purpose of testing, his rapport with examiner, and other factors
-

Source: Adapted from Robert L. Thorndike, *Personnel Selection* (New York: John Wiley and Sons, Inc., 1949), p. 73 and Lee Cronbach, "Test Reliability: Its Meaning and Determination," *Psychometrika*, Vol. 12 (January 1947), pp. 1-16.

"The term 'lasting' refers to consistency from one time to another, or in this context, from one testing occasion to another.

"The term 'general' refers to consistency from one sampling of stimuli or content to another; or in this context, consistency of examinee performance from one 'form' of the test to another equivalent form. When internal-consistency procedures are used to study reliability, the meaning of the word 'form' is stretched to include alternative combinations of items within a single test, such as odd-numbered items and even-numbered items.

mean (or the mean of the squared deviation scores). The relationship between the variance and the standard deviation is as follows:

$$V = \frac{\sum x^2}{N} = SD^2$$

$$SD = \sqrt{V}$$

In any set of test scores or other measures, the total variance (V) includes:

1. variance with respect to the ability or trait we are attempting to measure.
2. a certain amount of invalid variance (due to individual differences in test-wiseness, cheating, and other *systematic errors*)
3. a certain amount of error variance, due to *compensating errors*, that tend to average out if repeated samplings of test behavior are obtained.

A person sometimes scores "too high" or "too low" in a specific sample of behavior, but these inconsistencies average out over the long run. A useful analogy would be the tendency for a player's daily batting average to be

"too high" or "too low" to represent him fairly because of chance factors; while in cumulated batting averages these positive and negative "errors" tend to cancel out, resulting in less error variance and more reliable or consistent scores.

Error variance of the "compensating error" type arises from two major sources:

1. Instrument-centered errors resulting from ambiguities in test questions and directions and also from the fact that we test only a sample rather than a total universe of information and skills.
2. Errors resulting from temporal fluctuations in the individual examinee—variations from one testing occasion to another in his attitudes, speed of working, and other factors.

If we define an individual's "true score" as the average of an infinite number of testings with the same instrument,³ we can think of each person's score on a single testing as equal to his true score plus an error. (The expression "plus" is used in an algebraic sense with no implication that errors cannot decrease as well as augment a score).

$$\text{Then } X = X_{\text{true}} + X_{\text{error}}$$

where X_{true} represents the "true score," as defined. The average of an infinite number of testings for an individual would be X_{true} , since the sampling errors would tend to cancel each other out as considerable data were cumulated.

The total variance in obtained scores for a group of examinees would equal the sum of the "true variance" and the error variance.

$$\text{Total variance} = \text{true variance} + \text{error variance}$$

$$\text{or } SD^2_{\text{total}} = SD^2_{\text{true}} + SD^2_{\text{error}}$$

If it were feasible to test students one hundred or more times (without changing their performance through practice, boredom, or resistance), each individual's scores on these many different testings would be distributed according to a normal frequency distribution, with the mean score from all

³ The assumption is made that no learning takes place during such readministrations. Actually the term "true score" is not a good one since the term "true" implies absolute quality or validity in the measuring instrument. Buros suggests the term "asymptotic score" for the limiting value that would be approached as one cumulated evidence from a very large number of estimates. Those with background in mathematics, will be familiar with the term "asymptote" as a limiting value and will agree that "asymptotic score" would be a more appropriate term than "true score." Oscar K. Buros, "Schematization of Old and New Concepts of Test Reliability Based on Parametric Models," to appear in *20th Yearbook*, National Council on Measurement in Education. Ames, Iowa: The Council, 1964.

these testings being his "true score." The *SD* of this frequency distribution of scores on repeated testings would be called standard error⁴ or *SE*.

Once we had these standard errors, we could compute a reliability coefficient from this information. The reliability coefficient for a set of scores is defined as the *proportion of the total variance that is "true" or nonerror variance*.

Expressed in a formula,

$$\begin{aligned}\text{Rel. coeff.} &= \frac{\text{"true" variance}}{\text{total variance}} \\ &= \frac{\text{total variance} - \text{error variance}}{\text{total variance}} \\ &= 1 - \frac{\text{error variance}}{\text{total variance}} \\ \text{or Rel. coeff.} &= 1 - \frac{SE^2}{SD^2}\end{aligned}$$

Let us assume that our estimate of the standard error for a test (that is, the average standard deviation of repeated measurements for individual students) is 5 points. The error variance in this set of scores would then be SE^2 , or 25. Let us also assume that the standard deviation for the distribution of scores (for all students in the group) is 15; here the total variance would be 15^2 or 225. Then, substituting in the formula given above, we can obtain the reliability coefficient.

$$\text{Rel. coeff.} = 1 - \frac{\text{error variance}}{\text{total variance}} = 1 - \frac{25}{225} = 1 - .11 = .89$$

As the reader has undoubtedly concluded, it is not feasible to obtain the value for the error variance by retesting students one hundred times. Therefore, other methods are used to obtain the reliability coefficient. Then, when that coefficient is obtained, the *SE* (standard error) is computed from a variation of the formula given above.

$$\text{Since the Rel. coeff.} = 1 - \frac{SE^2}{SD^2}$$

we can obtain, by transposing terms:

$$\frac{SE^2}{SD^2} = 1 - \text{Rel. coeff.}$$

⁴ Actually standard errors can be computed for many types of statistics (for example, SE_M denotes the standard error of the mean, or the standard deviation of a distribution of means that would be obtained by taking an infinite number of samplings from a population). However, we will not use a subscript but will assume that *SE* refers to the standard error of measurement.

Then, by multiplying both sides of the equation by the denominator, we obtain:

$$SE^2 = SD^2 (1 - \text{Rel. coeff.})$$

Taking the square root of each term gives the typical formula for the standard error.

$$SE = SD \sqrt{1 - \text{Rel. coeff.}}$$

This formula is often written $SE = SD \sqrt{1 - r_{tt}}$ with r_{tt} representing the correlation coefficient between individuals' scores on one testing and their scores on another.

COMPUTING CORRELATION COEFFICIENTS

We must now turn our attention, from our discussion of reliability, to the correlation coefficient itself. The correlation coefficient is an exceedingly useful measure of relationship, which expresses in a single decimal fraction the degree of relationship or "going-togetherness" between two variables, such as the tendency for persons who are tall to make more "baskets" in a basketball game, or the tendency for students of higher IQ to have more extensive vocabularies.⁵

In many situations throughout this textbook, we will be concerned with relationships between two variables (for example, scores on two tests for the same students). Relationships between two sets of data are studied in order to answer the question: How well can I *predict* a person's relative status in one characteristic if I know his status in another characteristic? For example, we might like to know the answers to such questions as:

How confident can I be that students' IQ's would be about the same if they were retested next week with the same test or with another form of the test? On the basis of the students' intelligence quotients on this test, how well can I predict their success in certain school subjects?

⁵ The values of the correlation coefficient range in size from .00 to 1.00 (values of + 1.00 or - 1.00 represent a perfect positive or a perfect inverse relationship, which enables one to know a person's rank on one variable if one knows his rank on the other). An example of a negative correlation, or inverse relationship, would be that between typing speed and number of typing errors (since typists with a lower rate tend to make a larger number of errors although the relationship would be far from perfect). Correlation does *not* imply causation; for example, for a given individual, increasing his typing speed might actually increase, rather than decrease, his errors; however, among a fairly heterogeneous group of office workers, there would be a tendency for the most rapid typists to be those who made the fewest errors since both rapidity of typing and fewness of errors reflect a composite of aptitude and experience.

In other words, the chief interest of educators in correlation is the practical one of knowing how much dependence they can place upon certain types of data available to them in predicting other types of data that are helpful in their work.

The Spearman Rank-Difference Method of Computing a Correlation Coefficient

When one is computing a correlation coefficient for a small group, the most practical method is the Spearman Rank-Difference method. Moreover, when one has only rank-order data, this method is the preferred one. The steps in the method are shown in Table 3.2, which illustrates the procedure for finding r_{ho} , the rank-difference coefficient of correlation. Although this method is not suitable for computing reliability coefficients, for which a larger number of cases is needed, it is included here to help the student understand the meaning of a correlation coefficient.

Table 3.2
Computation of the Coefficient of Correlation
by the Rank-Difference Method

STUDENT	Grade Placements		Ranks ^a		<i>D</i>	<i>D</i> ²
	ARITH.	READING	ARITH.	READING		
	REASON- ING		REASON- ING			
1	5.5	6.2	8.	7.5	0.5	0.25
2	4.0	4.9	25.	19.	6.0	36.00
3	6.0	5.7	5.	12.	7.0	49.00
4	5.4	4.8	9.5	20.	10.5	110.25
5	4.8	5.5	16.	15.5	0.5	0.25
6	5.2	5.7	12.	12.	0.0	0.00
7	6.2	7.2	2.5	2.	0.5	0.25
8	4.7	3.8	20.	29.	9.0	81.00
9	4.8	5.0	16.	18.	2.0	4.00
10	4.6	4.7	22.	21.	1.0	1.00
11	4.8	5.6	16.	14.	2.0	4.00
12	5.2	4.1	12.	27.5	15.5	240.25
13	5.9	6.8	6.	3.	3.0	9.00
14	5.4	6.0	9.5	10.	0.5	0.25
15	6.8	7.9	1.	1.	0.0	0.00
16	6.2	6.2	2.5	7.5	5.0	25.00
17	3.9	4.4	27.	24.	3.0	9.00
18	4.8	5.4	16.	17.	1.0	1.00
19	4.5	4.1	23.	27.5	4.5	20.25
20	4.8	4.6	16.	22.5	6.5	42.25

Table 3.2 (Continued)
Computation of the Coefficient of Correlation
by the Rank-Difference Method

STUDENT	Grade Placements		Ranks ^a		<i>D</i>	<i>D</i> ²
	ARITH. REASON- ING	READING	ARITH. REASON- ING	READING		
21	5.2	6.5	12.	5.	7.0	49.00
22	4.0	6.4	25.	6.	19.0	361.00
23	6.1	5.7	4.	12.	8.0	64.00
24	4.0	4.6	25.	22.5	2.5	6.25
25	3.7	4.3	28.5	24.5	4.0	16.00
26	4.7	3.3	20.	30.	10.0	100.00
27	3.7	5.5	28.5	15.5	13.0	169.00
28	5.6	6.7	7.	4.	3.0	9.00
29	4.7	6.1	20.	9.	11.0	121.00
30	3.5	4.3	30.	24.5	5.5	30.25

 $N = 30$ $N^2 = 900$ $N^2 - 1 = 899$ $1558.50 = \Sigma D^2$ $9351 = 6(\Sigma D^2)$

$$\rho(rho) = 1 - \frac{6 \Sigma D^2}{N(N^2 - 1)}$$

$$\rho = 1 - \frac{6(1558.5)}{30(900 - 1)} = 1 - \frac{9351}{26970} = 1 - .347 = +.653.$$

^a When grade placements (or scores) are identical, their ranks are averaged. For example, in arithmetic reasoning, pupils 7 and 16 have identical grade placements of 6.2; hence ranks 2 and 3 are averaged; and a rank of 2.5 is assigned to each pupil.

An examination of the pairs of scores in the first two columns of Table 3.2 indicates a tendency for higher reading grade placements to be associated with higher grade placements in arithmetic reasoning. This relationship is more clearly evident when each student is assigned his rank in the group, first with respect to arithmetic reasoning, then with respect to reading. Comparison of these two columns of ranks reveals that a few students have identical ranks on the two tests; a large number of students have ranks that agree closely; whereas a small number show marked differences in rank. In the fifth column, the difference in ranks (*D*) is shown for each student. In the last column, the *D* value for each student has been squared. The sum of the last column (ΣD^2) is 1558.5. It can readily be seen that the closer the agreement between ranks, the smaller this sum,

and the larger the correlation. At the bottom of the table, the steps in the computation of *rho* are shown. How to interpret the correlation coefficient of .653 will be explained in a later chapter section.

The Pearson Product-moment Method of Computing Correlation Coefficients

There are several different methods of computing coefficients of correlation, which can be studied in a standard textbook on statistics. Each has its special uses. It is beyond the scope of this book to attempt an explanation of all these methods. In research and test construction, the method of computation most frequently used is the Pearson product-moment method, which we will now consider.⁶

THE Z-SCORE FORMULA FOR PEARSON *r* There are a number of equivalent formulas for computing the coefficient of correlation, or *r*, by the Pearson product-moment method. The formula given below can be used only when the data for both the *x* and *y* variables have been changed into *z*-scores.

$$r = \frac{\sum z_x z_y}{N}$$

Where z_x stands for the standard scores in the *x* variable (reading) and z_y stands for the standard scores in the *y* variable (arithmetic reasoning) and $\sum z_x z_y$ represents the sum of the products of pairs of standard scores.

In Table 3.3 this formula is applied to the same data previously used in the computation of *rho*. This standard-score formula is seldom used in practice, because of the work involved in computing numerous standard scores. However, this formula does illustrate the basic principles underlying the Pearson product-moment method. The product of *each pair* of standard scores for each student is first obtained. The value of *r*, as the *average of all these products*, is then computed.

It may be difficult to see why the average of these products is a good measure of relationship. The examples given below help one to understand.

1. Students 7 and 15 have high positive deviations from the means of both *x* and *y*; they illustrate a close relationship between arithmetic *GP* and reading *GP*; their products are large and therefore increase the value of *r* (the average product).
2. Students 25 and 30 show high negative deviations from the means of both *x* and *y*; they too illustrate a close relationship between arithmetic *GP* and *RGP*; and since they have large products, increase the value of *r*.

⁶ The use of this method assumes a linear relationship between the two variables; that is, we assume that a straight line can be drawn that will represent reasonably well the tallies for pairs of scores (as shown in the scatter-diagram of Figure A.2) in Appendix D.

Table 3.3
Computation of the Coefficient of Correlation
by the Pearson Product-Moment Method
 (Standard-score or z-score formula: $r = \frac{\sum z_x z_y}{N}$).

STUDENT	Grade Placements ^a		Standard or z-Scores ^b		
	ARITH.		ARITH.		
	REASONING	READING	REASONING	READING	PRODUCT
			z_y	z_x	$z_x z_y$
1	5.5	6.2	+ .6	+ .7	+ .42
2	4.0	4.9	-1.3	- .5	+ .65
3	6.0	5.7	+1.3	+ .3	+ .39
4	5.4	4.8	+ .5	- .5	- .25
5	4.8	5.5	- .3	+ .1	- .03
6	5.2	5.7	+ .3	+ .3	+ .09
7	6.2	7.2	+1.5	+1.6	+2.40
8	4.7	3.8	- .4	-1.5	+ .60
9	4.8	5.0	- .3	- .4	+ .12
10	4.6	4.7	- .5	- .6	+ .30
11	4.8	5.6	- .3	+ .2	- .06
12	5.2	4.1	+ .3	-1.2	- .36
13	5.9	6.8	+1.1	+1.3	+1.43
14	5.4	6.0	+ .5	+ .5	+ .25
15	6.8	7.9	+2.3	+2.3	+5.29
16	6.2	6.2	+1.5	+ .7	+1.05
17	3.9	4.4	-1.4	- .9	+1.26
18	4.8	5.4	- .3	0.0	0.00
19	4.5	4.1	- .7	-1.2	+ .84
20	4.8	4.6	- .3	- .7	+ .21
21	5.2	6.5	+ .3	+1.0	+ .30
22	4.0	6.4	-1.3	+ .9	-1.17
23	6.1	5.7	+1.4	+ .3	+ .42
24	4.0	4.6	-1.3	- .7	+ .91
25	3.7	4.3	-1.6	-1.0	+1.60
26	4.7	3.3	- .4	-1.9	+ .76
27	3.7	5.5	-1.6	+ .1	- .16
28	5.6	6.7	+ .8	+1.2	+ .96
29	4.7	6.1	- .4	+ .6	- .24
30	3.5	4.3	-1.9	-1.0	+1.90
$r = \frac{\sum z_x z_y}{N} = \frac{19.88}{30} = + .662.$					$\sum z_x z_y = 19.88$

^a Reproduced from Table 3.2.

^b The z_y , or standard-score values, in the fourth column were obtained by translating each arithmetic reasoning GP into a standard score by use of the formula $z = \frac{X - M}{SD}$, in which X stands for the original grade placement, M for the mean GP of 5.0, and SD for the standard deviation of 0.8. The z_x , or standard-score values, in the fifth column were obtained by translating each RGP value into a standard score by means of the same formula, in which X stands for the original RGP, M for the mean RGP of 5.4, and SD for the standard deviation for RGP's of 1.1.

3. A pair of scores (as for student 22) with an arithmetic *GP* markedly *below* average and an *RGP* definitely *above* average illustrate a negative or inverse relationship between the two variables; the product of standard scores in this case is negative and *decreases* the value of r .

TABULATING DATA IN A SCATTER-DIAGRAM Ordinarily, the first step in the Pearson method is to tabulate pairs of scores in a *scatter-diagram*, similar to those in Figure 3.1. Class intervals are set up according to the same principles used in setting up intervals for frequency distributions. In preparing a scatter-diagram for use in correlation, a tally is entered for each pair of individual scores, and the number of tallies in each square or cell is totaled.

In preparing Figure A.2 in Appendix D, 500 pairs of scores were tallied to obtain a reliability coefficient by the subdivided-test or split-halves method.⁷ That is, for each student a tally was entered in the square corresponding to his score on the even-numbered questions (the y -variable) and his score on the odd-numbered questions (the x -variable). Scores are from the 100-item state history test.

Interpretation of Correlation Coefficients

Before we interpret reliability coefficients, we will consider the more general problem of what a correlation coefficient means in terms of accuracy of prediction.

INTERPRETATION OF r IN TERMS OF SLOPE OF THE PREDICTION LINE AND THE AMOUNT OF REGRESSION OF PREDICTED SCORES TOWARD THE MEAN A study of the relationship between two variables is usually made to ascertain how accurately we can predict students' scores on one variable from our knowledge of their scores on another. For example, we might predict a student's score on an algebra test from his score on a test in arithmetic. If there were no relationship between the two variables and we had no information to go on, our safest guess would be to predict that each student would make an average algebra score. In such a situation, information about arithmetic test scores would be of no help. On the other hand, if there were a perfect correlation of 1.00 between these two variables, the error of prediction would be reduced to zero, and each student's rank on the algebra test would be the same as his rank on the arithmetic test.

Our prediction equation (in z -score form) is:

$$\bar{z}_y = r z_x$$

with \bar{z}_y standing for the predicted standard score in the y variable. When r is 0, $\bar{z}_y = (0) z_x$; the best prediction for any student is a z -score of 0, or

⁷ This method is explained on page 86.

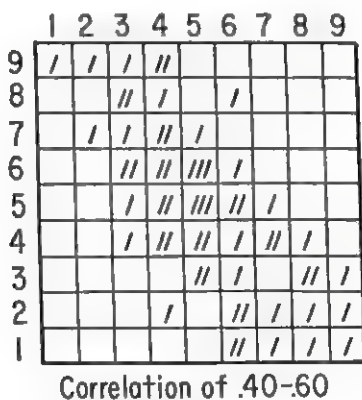
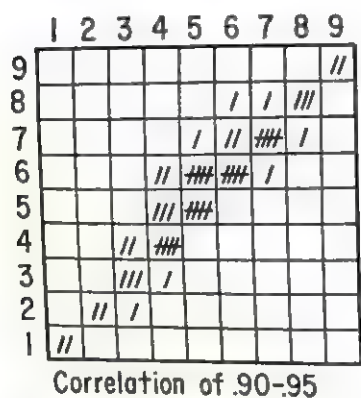
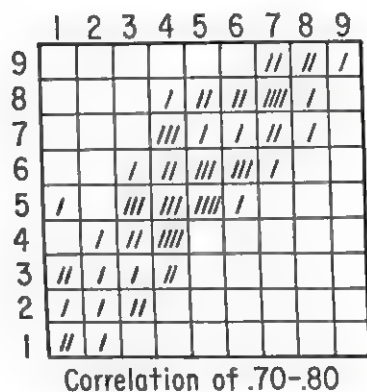
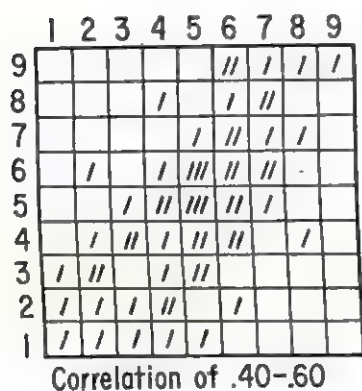
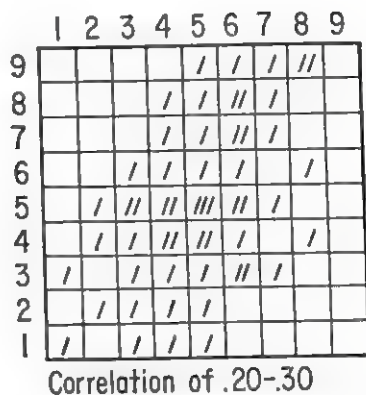
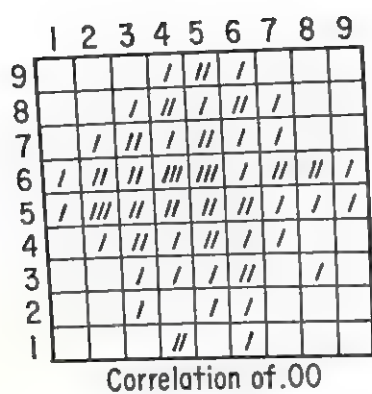


Fig. 3.1 Illustrative Scatter-Diagram Yielding Correlation Coefficients of Various Sizes.

M_x . If r is 1.00, $\bar{z}_y = (1) z_x$, and the predicted \bar{z}_y for each student is the same as his z_x .

Mathematics majors will realize that r , or $\frac{\bar{z}_y}{z_x}$, is the slope of the line of "best fit" (that is, the line that most accurately summarizes the relationship between the two variables). The equation of this prediction line ($\bar{z}_y = r z_x$) can be used in predicting the value of \bar{z}_y for any student if we know his z_x . The closer the relationship between x and y , the closer the line of "best fit" approaches a 45° angle with the x -axis, and the closer r (the slope of the prediction line) approaches 1.00.

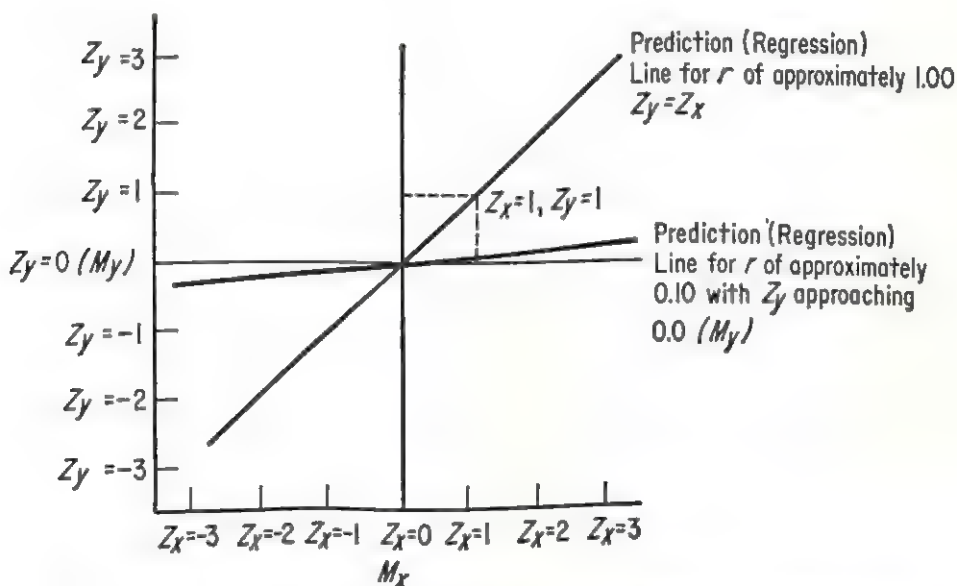


Fig. 3.2 Prediction (Regression) Lines for Perfect Correlation and Extremely Low Correlation.

If we tallied all pairs of z -scores, we could draw this line of best fit by first plotting the average scores of the columns, and then drawing a line by visual inspection that best fits this set of points.⁸ The slope of this prediction line would give us a rough estimate of the value of r (Fig. 3.2).

The lower the relationship between x and y , the closer the line of best fit approaches a 0° angle with the x axis, and the closer r (the slope of the prediction line) approaches 0. Obviously, r is the slope of the predic-

⁸ If the averages lie on a curve, rather than a straight line, a different method of computing the correlation coefficient should be used since Pearson r would underestimate the degree of relationship.

tion line only if the x and y variables are in comparable units, for example, z -scores.

We could substitute any other value of r that we might obtain in this prediction equation; for example, if r is .70, $z_y = .7 z_x$, while if r is only .30, $\bar{z}_y = .3 z_x$. One can see why the prediction line has come to be known as the regression line, since r is a measure of the extent to which predicted scores *regress* toward the mean; the lower the correlation, the greater the regression.

INTERPRETATION OF r^2 IN TERMS OF PERCENT OF VARIANCE EXPLAINED AND THE STANDARD ERROR OF ESTIMATE On the scatter-diagrams of Figure 3.1 which yield a high r , the tallies cluster closely around the prediction line, the small amount of scatter of tallies around the prediction line indicating a closer relationship and relatively greater accuracy of prediction. When r is very low, the tallies are scattered throughout the columns in such a way that it is obvious that no sound basis exists for predicting y from x .

If we could obtain an approximate measure of the amount of scatter or variability of scores around this prediction line, it would represent our error of prediction or estimate. We could compute the SD for each column, which reflects the undependability of predicting y from x ; then we could obtain a weighted average⁹ of these column SD 's. This average would be a fair approximation of the standard error of predicting \hat{y} from x ($SE_{y \cdot x}$).

The higher the ratio of the "error of prediction" variance ($SE_{y \cdot x}^2$) to the total variance (SD_y^2), the lower the relationship between the two variables. Conversely, the lower the ratio of $SE_{y \cdot x}^2$ to SD_y^2 , the closer the relationship. $SE_{y \cdot x}^2$ represents the "unpredictable" variance.

$$\begin{array}{l} \text{Proportion of} \\ \text{"unpredictable"} \\ \text{variance} \end{array} = \frac{SE_{y \cdot x}^2}{SD_y^2}$$

But r^2 represents the proportion of variance in one variable that is predictable from the other. Therefore

$$\begin{aligned} r^2 &= 1 - \text{proportion of unpredictable variance} \\ &= 1 - \frac{SE_{y \cdot x}^2}{SD_y^2} \end{aligned}$$

If we wished, we could estimate r by (1) computing the proportion of variance that is "unpredictable" or due to errors of estimate, and then (2) subtracting this ratio from 1.00 to obtain the proportion of predictable variance or r^2 . Actually, r is usually computed by the Pearson product-

⁹ Weighted in terms of the number of cases in each column.

moment method and the standard error of estimate is obtained by the following variation of the formula given above:

$$SE_{y \cdot x} = SD_y \sqrt{1 - r^2}$$

When r is .93 (as in Figure A.2), $r^2 = .86$. Hence, we can say that 86 percent of the variance in the even-numbered scores is predictable from knowledge of the odd-numbered scores. If all examinees had identical odd-numbered scores, the variance in the even-numbered scores would be reduced by 86 percent; only 14 percent of the variance, attributable to other factors, would remain.

After this long "detour" to consider the computation and interpretation of correlation coefficients, we can now return to the subject of reliability. The student should review the concept of, and the formula for, the standard error (SE) of a test score before studying the next chapter section.

COMPARISON OF STANDARD ERRORS AND RELIABILITY COEFFICIENTS AS MEASURES OF RELIABILITY

The concept of reliability is perhaps most easily understood in terms of the standard error. One can readily see how the standard deviation of a series of scores on repeated tests for the same individual reflects (1) the consistency of the test as a measuring instrument and (2) the fluctuations with time in certain characteristics of the examinee that affect his test scores. Use of the standard error also helps us to think of a student's test score as representing a range of probable scores. That is, if the SE for a specific intelligence test is 5 points, we can infer that the chances are two out of three that scores obtained by students differ from their "true scores" by less than 5 points. These "odds" are based on the fact that two-thirds of the scores lie within one SD of the mean (in this case, the individual's theoretical true score).

Although the standard error is quite valuable in the interpretation of individual test scores, reliability coefficients are preferred for comparing the consistency of measurement of *different tests* (for example, our locally developed arithmetic, spelling, and history tests). Standard errors are not comparable from one test to another; that is, the SE varies in size with the number of items on the test, as well as the SD of the test scores. When one wishes to compare the reliability of two or more tests, the reliability coefficients usually provide a better basis for comparison than do the standard errors.¹⁰

¹⁰ However, the standard errors of two tests both utilizing T -scaled scores (or some other type of standard score) would provide just as good a basis as the reliability coefficients for comparing the tests with respect to consistency of measurement. When T -scaled scores are used, the SD 's for both tests would be 10; and the SE 's would therefore be comparable.

METHODS OF ESTIMATING RELIABILITY OF TEST SCORES

Reliability coefficients can be obtained in several different ways. When one examines Table 3.4, one realizes that there can be no single reliability coefficient for a test. Moreover, since standard errors are based on reliability coefficients, there is similarly no such thing as *the* standard error for a test. A standard error reflects the types of error variance that are measured by the reliability coefficient on which it is based. The four most frequently used approaches will be considered in turn.

Table 3.4
Comparison of Different Types of Reliability Coefficients with
Respect to Types of Error Variance Taken into Account

TYPE OF COEFFICIENT	METHOD USED	TYPES OF SCORE VARIANCE ^a COUNTED AS	
		True variance	Error variance
TEST-RETEST METHOD			
Coefficient of stability (consistency over time on same content)	Same test sample	LG, LS	TG, TS
	Different occasions	Lasting general and lasting specific	Temporary general and temporary specific
EQUIVALENT FORMS METHOD			
Coefficient of equivalence (consistency in performance on specific content samples)	Same occasion	LG, TG	LS, TS
	Different test samples (that is, parallel forms administered at essentially the same time)	Lasting general and temporary general	Lasting specific and temporary specific
INTERNAL-CONSISTENCY METHODS			
Coefficient of internal consistency (approximation of coefficient of equivalence)	Internal analysis of data (same occasion, different samplings of items from same test; that is, subdivided-test or Kuder-Richardson method) ^b	LG, TG	LS, TS
		Lasting general and temporary general	Lasting specific and temporary specific

TYPE OF COEFFICIENT	METHOD USED	TYPES OF SCORE VARIANCE ^a COUNTED AS	
		True variance	Error variance
EQUIVALENT-FORMS METHOD WITH TIME INTERVAL			
Coefficient of equivalence and stability (consist- ency over time and over specific content samples)	Different occasions	LG	LS, TG, TS
	Different tests, i.e. administration of parallel forms with intervening time interval)	Lasting general	Lasting specific, temporary gen- eral, and tempo- rary specific

^a For examples of factors contributing to LG (lasting general), LS (lasting specific), TS (temporary specific), and TG (temporary general), the reader is referred to Table 3.1.

^b Although the Kuder-Richardson formula measures consistency of examinee performance on all items in the test, use of the Kuder-Richardson formula 20 results in a reliability coefficient which approximates the average of all the split-half coefficients which would be obtained on all possible divisions of the test into equivalent halves. The Kuder-Richardson 20 formula is computed from data concerning the proportion of examinees passing each test item and the SD of test scores. See J. P. Guilford, *Fundamental Statistics in Psychology and Education* (New York: McGraw-Hill Book Company, Inc., 1956), pp. 454-455.

The Test-Retest Method

When reliability is measured by the test-retest method, a *coefficient of stability* is obtained. This reliability coefficient measures error variance due to *temporal* variations in characteristics of the examinee, as well as variation in conditions of test administration. Some of this temporal instability in test scores is due to variations from one testing occasion to another in the examinees' *general* characteristics, such as in his health or emotional tension; part of it is due to variations in their reactions to the *specific* test. Illustrations of these sources of variance are listed in Table 3.1 under the headings TG and TS respectively. In other words, when the test-retest method is used, a coefficient of stability is obtained, which reflects only the TG and TS types of error variance (that is, variations in examinee test performance from one testing occasion to another).

When the test-retest method is used, the interval between tests should be at least several days so that the student's memory of his answers does not spuriously increase the consistency of scores. However, the time interval should not exceed two or three weeks because we are trying to measure stability of student performance on the test, rather than the stability of the interest, ability, or personality trait measured.¹¹

¹¹ A correlation coefficient, based on two testings between which opportunity for learning and/or maturation has occurred, is a useful statistic, especially for a

The test-retest method is infrequently used. Practice effects, which are not the same for all subjects, interfere with our attempt to measure the test's consistency. Moreover, students may be unwilling or unable to retake the test with the same level of motivation. Perhaps even more serious is the fact that the test-retest method fails to measure the types of error variance listed under LS, which result from the fact that a specific test includes only a sampling of content from the area that the test is designed to represent.

The Equivalent-Forms Method

Many standardized tests have two or more equivalent forms that have been designed to be comparable in content, length, difficulty level, and variance. When two equivalent forms (say forms A and B) are administered to students on the same occasion,¹² a *coefficient of equivalence* is obtained, which measures the consistency of examinee performance from one specific sampling of test content to another. With this procedure, error variance due to TS and LS are measured, that is, temporary and lasting characteristics of the examinee that are elicited by this specific sampling of test items. This method does not take into account temporal fluctuations in examinee performance.

The Internal-Consistency Methods

Sometimes equivalent forms of a test are not available; sometimes it is difficult to obtain permission to have students take both forms of a test. For these reasons, internal consistency methods of measuring reliability have become popular. Such methods involve a comparison of examinee performance on different samplings of items *from the same test*. They provide satisfactory estimates of the coefficient of equivalence.

A frequently used "internal consistency" method of estimating test reliability is the subdivided-test method, often called the split-halves method or the odd-even method. The last name arises from the fact that for most tests, the efficient way of computing such a reliability coefficient is to score the odd-numbered items as one "form," the even-numbered items as an-

counselor who is using present test data on aptitudes or interests as a basis for inferences about future performance. However, such a coefficient should not be interpreted as simply estimating reliability, or error variance; it requires a more complex interpretation.

¹² In order to equate the effects of practice on student achievement on the two forms, it is good procedure to have half the students take form A and then form B; and to have the other half take form B first, followed by form A.

other "form," and then to correlate students' scores on the two halves of the test.¹³

The subdivided-test method, like the equivalent-forms method, takes into account variance due to the specificity of the tests and fails to measure temporal instability in test performance of students. In Figure A.2, the reliability coefficient has been computed by this method for the state history test. The reliability coefficient is .93. Correction with the Spearman-Brown formula would give a corrected coefficient¹⁴ of .96.

The Kuder-Richardson method, like the subdivided-test method, is based on consistency in the student's test performance on different items. However, while the subdivided-test method compared students' scores on two halves of the test, the Kuder-Richardson method involves a study of inter-item consistency. With a relatively simple approximation formula, known as the Kuder-Richardson Formula 21, reliability coefficients can be computed quite easily, just on the basis of the mean (M), the number of items (n) and the standard deviation (SD)

$$r_{KR\ 21} = \frac{n}{n-1} \left(1 - \frac{M(n-M)}{n(SD)^2} \right)$$

For example, on the state history test, where n is 100, the mean is 78, and the SD is approximately 10, the computation is as follows:

$$\begin{aligned} r_{KR\ 21} &= \frac{100}{99} \left(1 - \frac{78 \times (22)}{100 \times (10)^2} \right) \\ &= 1.01 \left(1 - \frac{1716}{10,000} \right) = 1.01 (1 - .17) \\ &= 1.01 (.83) = .84 \end{aligned}$$

¹³ It would not be satisfactory to score items in the first half of the test as one "form," with the remaining items constituting the other "form." Because easier items are included in the first half of the test, such a procedure would not result in "forms" of equal difficulty. Moreover, they would be dissimilar with respect to content.

¹⁴ When the reliability coefficient is computed in this way, a correction must be made for the fact that each of these halves is only half as long as the original test. Since a larger sampling of items results in greater consistency of measurement, the Spearman-Brown correction is applied to estimate the reliability coefficient that would be obtained for a test twice the length of the odd or even-numbered "form." The table for Spearman-Brown corrections is in Appendix D (Table A.2). However, the following formula is superior in that it is not based on the assumption (made in deriving the Spearman-Brown formula) that the SD 's of the two half-scores are equal:

$$r_{tt} = 2 \left(1 - \frac{SD_a^2 + SD_b^2}{SD_t^2} \right)$$

where SD_a and SD_b are the standard deviations of scores on the two half-tests and SD_t is the standard deviation of scores on the total test.

If the longer Kuder-Richardson Formula 20 is used, a reliability coefficient is obtained that approximates the average of all split-half coefficients which would be obtained on all possible divisions of the test into equivalent halves.¹⁵ For most tests, however, the simpler Formula 21 will give nearly the same results.

For a still easier method of approximating a Kuder-Richardson reliability coefficient, the reader is referred to Table 3.5. To illustrate the use of this table, we will determine the reliability of the state history test. We need to know only that the test contains 100 items, that the average score is 78 percent, and that the *SD* is $\frac{1}{10}$ as large as the number of items, or equal to $.10n$. Using the first table (A) for comparatively easy tests, we obtain a reliability coefficient of .85, which agrees very closely with the coefficient computed by Formula 21.

From this same table we could also obtain the Kuder-Richardson reliability coefficient for the arithmetic test from the data given in Table 2.3. This test is also an easy test, with a *M* of 84 out of 100 items. The reliability coefficient would be approximately .85 if the *SD* were $.10n$ and .94 if the *SD* were $.15n$. Since the *SD* is $.11n$, or one-fifth the difference between the two values heading the columns, we can interpolate and obtain a reliability coefficient of $.85 + \frac{1}{5} (.10)$ or $.85 + .02$ or .87. Thus, we see that these two locally developed tests are almost equally reliable when their reliability is measured by the Kuder-Richardson method.

Table 3.5
Approximate Kuder-Richardson Reliability Coefficients
A. Reliability Coefficients for Comparatively Easy Tests
(Average Score—70% to 90% correct)

NO. OF ITEMS (<i>n</i>)	Reliability Coefficient When <i>SD</i> Is		
	$.10n$	$.15n$	$.20n$
100	.85	.94	.97
90	.83	.93	.97
80	.81	.92	.96
70	.78	.91	.96
60	.75	.90	.95
50	.69	.88	.94
40	.62	.84	.92
30	.48	.80	.90
20	.21	.68	.84

¹⁵ This formula requires the computation of the proportion of students passing and failing each item.

B. Reliability Coefficients for Comparatively Difficult Tests
(Average Score—50% to 70% correct)

NO. OF ITEMS (<i>n</i>)	Reliability Coefficient When <i>SD</i> Is		
	.10 <i>n</i>	.15 <i>n</i>	.20 <i>n</i>
100	.77	.90	.95
90	.74	.89	.94
80	.71	.88	.94
70	.66	.86	.93
60	.61	.84	.92
50	.53	.80	.90
40	.41	.75	.87
30	.21	.67	.83
20		.49	.74

Source: Adapted from tables in Paul Diederich, *Short-Cut Statistics for Teacher-Made Tests*, Evaluation and Advisory Service Series No. 5 (Princeton, N. J.: Educational Testing Service, 1960), p. 29.

Table 3.6
Standard Error of Measurement for Different Values of the Reliability Coefficient and the Standard Deviation of Test Scores^a

RELIABILITY COEFFICIENT	SE When Standard Deviation ^b Is									
	1	2	3	4	5	6	7	8	9	10
.98	0.1	0.3	0.4	0.6	0.7	0.8	1.0	1.1	1.3	1.4
.95	0.2	0.4	0.7	0.9	1.1	1.3	1.6	1.8	2.0	2.2
.90	0.3	0.6	0.9	1.3	1.6	1.9	2.2	2.5	2.8	3.2
.85	0.4	0.8	1.1	1.5	1.9	2.3	2.7	3.1	3.5	3.9
.80	0.4	0.9	1.3	1.8	2.2	2.7	3.1	3.6	4.0	4.5
.75	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
.70	0.5	1.1	1.6	2.2	2.7	3.3	3.8	4.4	4.9	5.5
.65	0.6	1.2	1.8	2.4	3.0	3.6	4.1	4.7	5.3	6.0
.60	0.6	1.3	1.9	2.5	3.2	3.8	4.4	5.1	5.7	6.3

^a The standard error of measurement is computed by multiplying the standard deviation of test scores by the radical $\sqrt{1 - \text{rel. coefficient}}$. In other words

$$SE = SD \sqrt{1 - r_{tt}}$$

^b The column headed 10 may be used whenever *T*-scores are involved. If the standard deviation of test scores is higher than 10, for example, 12 (as in the state history test), one can add the appropriate entries in the columns for standard deviations of 10 and 2 (or other appropriate combinations). In this case the *SD* is 12, the estimated reliability coefficient (from Table 3.5) is .85; hence the *SE* is $3.9 \div .8$ or 4.7. There are two chances in three that a student's obtained score in the state history test differs from his "true score" by not more than 4.7 points. A student's "true score" is a theoretical score—the average of all scores obtained on an infinite number of retestings.

By the use of Table 3.6, we find that for tests with a reliability coefficient of approximately .85, the standard error of a T -score for both the arithmetic and history tests ($SD = 10$) is 3.9, or approximately four score points.

The Method Involving Equivalent Forms Administered with a Time Interval between Testings

The reader will note from Table 3.4 that the administration of equivalent forms, with an interval of time between the two administrations, measures all the major types of error variance. Certainly, if the test user intends to measure student *growth* during some time interval by administering equivalent forms of a test on different occasions, this type of reliability coefficient that measures *both equivalence and stability* is the one which should be computed. Naturally, reliability coefficients computed by this method will tend to be lower than those computed by any other method because more types of error variance are taken into account.

Comparison of Methods

Some measure of the stability of scores from one testing occasion to another should be given in any test manual. Ideally, the last method (which yields a coefficient of equivalence and stability) should be used; this method minimizes recall of specific answers and avoids other difficulties mentioned under the discussion of the test-retest method.

The internal consistency methods help one to estimate how much of the variance in test scores is due to lack of equivalence between different samplings of items. The Kuder-Richardson method is better than the subdivided-test method in that it provides an estimate of the average coefficient that could be obtained if all possible subdivisions of the test were utilized.

The Kuder-Richardson method measures equivalence through a study of interitem consistency in student performance. If the universe of items we are sampling is fairly homogeneous, student performance will be fairly consistent from item to item. If the universe is very homogeneous, interitem consistency will be unusually high; only a relatively small sample will be needed as a basis for inferences about student performance on the universe of possible items.

This relationship of reliability to homogeneity of content is analogous to the situation we find in laboratory analysis of blood where a very small sample of the homogeneous "universe" being measured serves as a reliable basis for many important inferences. No dentist, however, would make inferences about the condition of one's teeth from the examination of two

or three teeth, even though such a sampling constitutes a much larger proportion of the whole than the sampling of blood that the laboratory technician obtained. The more homogeneous the content sampled, the more consistent the results from sample to sample.

Table 3.7
Reliability Coefficients for Four Subtests of the SRA Primary
Mental Abilities for Ages 11 to 17

SUBTEST	Reliability Coefficient Found by	
	SPLIT-HALF METHOD	SEPARATELY-TIMED HALVES
Verbal Reasoning	.94	.90
Reasoning	.96	.87
Space	.90	.75
Number	.92	.83

Source: Anne Anastasi and J. D. Drake, "An Empirical Comparison of Certain Techniques for Estimating the Reliability of Speeded Tests," *Educational and Psychological Measurement*, vol. 14 (Autumn 1954), pp. 529-540.

Neither the Kuder-Richardson nor the subdivided-test method should be used if tests are highly speeded.¹⁶ The way in which an internal-consistency method can spuriously inflate the reliability coefficients of speeded tests is shown in Table 3.7. The authors estimated the extent to which each subtest of the *Primary Mental Abilities Test* was speeded.¹⁷ Their findings indicated that Verbal Reasoning scores were least affected by individual differences in working speed; that scores in the Reasoning test were somewhat affected by speed; and that the Space and Number tests were highly speeded. As Table 3.7 shows, the reliability coefficients of the

¹⁶ On a highly speeded test, students would tend to get similar scores on chance-halves of the test simply because their speed of working in a specific testing session would have enabled them to cover numbers of items on the two halves of the test. If the odd-numbered items and even-numbered items were typed or printed as separate forms and administered under separate time limits, the split-halves method would be acceptable. A test should be considered a speeded test if there is considerable variation with respect to the number of items omitted at the end of the test, or if there is a low correlation between scores earned on the test when administered with and without time limits.

¹⁷ The method used was to find the variance with respect to number of items completed by students and then divide this variance by the total variance of test scores.

relatively unspeeeded verbal reasoning test were approximately the same when the conventional split-halves method was used, as compared with the separately-timed-halves method. For the other tests, the conventional split-halves method gave an inflated estimate of reliability.

RELIABILITY OF DIFFERENCE SCORES

When we compare the scores of students in two tests (for example, the arithmetic and history tests), we wish to know whether the differences are largely attributable to errors of measurement or whether we would be likely to find similar intraindividual differences if we retested. That is, we are concerned with the reliability of "difference scores." Unfortunately, the reliability of differences between pairs of scores is much less than the reliability of either score. Two factors are responsible for the lower reliability of difference scores: (1) the errors of measurement in both tests affect the error variance of the difference; and (2) whatever is common to both tests (arithmetic and state history) is canceled out in the com-

Table 3.8
Reliability of Differences^a between Standard Scores

CORRELATION BETWEEN TWO TESTS	RELIABILITY COEFFICIENT FOR DIFFERENCE SCORES WHEN AVERAGE RELIABILITY COEFFICIENT OF TWO TESTS IS					
	.70	.75	.80	.85	.90	.95
.00	.70	.75	.80	.85	.90	.95
.10	.67	.72	.78	.83	.89	.94
.20	.63	.69	.75	.81	.88	.94
.30	.57	.64	.71	.79	.86	.93
.40	.50	.58	.67	.75	.83	.91
.50	.40	.50	.60	.70	.80	.90
.60	.25	.38	.50	.62	.75	.88
.70	.00	.17	.33	.50	.67	.83
.75		.00	.20	.40	.60	.80
.80			.00	.25	.50	.75
.85				.00	.33	.67
.90					.00	.50
.95						.00

^a Computations are based on the following formula

$$\text{Rel. diff.} = \frac{\text{Av. rel. coef. of 2 tests} - \text{Intercorrelation between 2 tests}}{1 - \text{Intercorrelation}}$$

putation of the difference.¹⁸ If two tests overlap considerably with respect to abilities measured, a considerable portion of the consistent variance in each score is due to the overlapping part. When that variance is subtracted, as it is in obtaining difference scores, the remaining test variance contains a larger *proportion* of error.

Let us assume that the intercorrelation between the arithmetic and history tests is .60. We can look in the appropriate row of Table 3.8 and find, in the column headed .85 (the average reliability coefficient of the two tests), that the reliability of the difference scores will be only .62.

If we wish to compute the standard error for the difference scores (for comparisons between the arithmetic and history tests, each of which has an *SE* of 4), we can use the following approximation formula.

$$SE_{diff} = \sqrt{SE_1^2 + SE_2^2} = \sqrt{4^2 + 4^2} = \sqrt{32} = 5.7$$

With a SE_{diff} of 6 points, the chances would be two out of three that a student's difference score was within 6 points of his true difference score.

FACTORS AFFECTING THE SIZE OF RELIABILITY COEFFICIENTS

Obviously, we would like test scores to be as reliable as possible. However, it would be a mistake to assume that we should simply list the reliability coefficients for several published tests, among which we wish to choose, and select the test with the highest coefficient as best for our purposes.

Methods of Estimating Reliability

Comparing the reliability coefficients for different tests requires careful attention to factors that affect the size of such coefficients. We have already considered one of the major factors affecting their size, that is, the *method* used in obtaining data on reliability. The method that takes into account both stability and equivalence will tend to give lower coefficients than the other methods because all major types of error variance are included. The Kuder-Richardson method will tend to yield lower coefficients than the split-halves method because the former reflects test homogeneity, as reflected in all interitem relationships, rather than merely the consistency of scores on two halves of a subdivided test.

With any method involving two testing occasions, the longer the interval of time between two test administrations, the lower the coefficient will tend to be.

¹⁸ If the correlation between two tests were positive and perfect, all the difference scores would be zero; if the correlation were -1.00 , the difference scores would be of maximum size. In between these two extremes, the larger the positive correlation between the tests, the smaller the differences will tend to be.

Heterogeneity of Group

Another factor affecting size of reliability coefficients is more difficult to understand, that is the "range of ability" or dispersion of scores within the group on which the reliability coefficient is computed. For example, the reliability coefficients for intelligence tests, computed on elementary or high school groups, tend to be higher than those computed on college groups, which show much less dispersion with respect to IQ.

We have already cited two different reliability coefficients for the state history test. The coefficient of .85 was obtained by the Kuder-Richardson method, the one of .96 by the split-halves method.¹⁹ This large difference cannot be explained entirely in terms of method. Another reason is that the reliability coefficient computed by the split-halves method was based on students from *all schools* in the district, a more heterogeneous population than the Central High School group, on which the Kuder-Richardson coefficient was obtained. The *SD* for "all schools" was 12, rather than 10. If the split-halves reliability coefficient had been computed on a sample with the smaller *SD* of 10, the coefficient would have been reduced²⁰ from .93 to .85. Hence, we see that some of the difference is due to method and some to differences in the homogeneity of the group studied.

Ideally, reliability coefficients should be computed on groups which have about the same *SD* as the group for which the test user wishes to make interindividual comparisons. The user of an achievement test usually wishes to compare individuals within a single grade in a school or a school district, rather than individuals within a larger, more heterogeneous population. Therefore, an increasing number of test manuals are presenting reliability coefficients by grade level for each of several schools, or each of several communities.

Table 3.9 illustrates good practice in that reliability coefficients were computed for single communities; the range of coefficients is shown, rather than coefficients based on the more heterogeneous population of all four school systems combined.

This table, however, like those in many test manuals, provides inadequate data concerning the samples of students tested. The Technical Recommendations specify that reliability samples should be described.²¹

¹⁹ The coefficient of .93, computed in Figure A.2, was corrected by the Spearman-Brown method (Table A.2) to .96.

²⁰ A table for correcting values of *r* for "restriction in range" is given in most standard textbooks in statistics, for example, Quinn McNemar, *Psychological Statistics*, 3d ed. (New York: John Wiley and Sons, Inc., 1962), p. 144.

²¹ "Technical Recommendations for Psychological Tests and Diagnostic Techniques" Supplement to the *Psychological Bulletin*, vol. 51 (March 1954), p. 230.

Table 3.9
Reliability Coefficients and Standard Errors of Measurement for
Subtests of *Metropolitan Achievement Tests*, Intermediate Level

TEST	r_{11}^a		$SE^b_{\text{Mens.}}$	
	RANGE	MEDIAN	RANGE	MEDIAN
1. Word Knowledge	.88-.95	.94	3.0-3.4	3.1
2. Reading	.89-.92	.90	2.5-2.8	2.6
3. Spelling	.91-.96	.92	2.6-3.5	3.0
4. Language				
Part A—Usage	.78-.84	.81	1.9-2.5	2.2
Part B—Parts of Speech	.64-.77	.72	1.3-1.3	1.3
Part C—Punctuation and Capitalization	.80-.88	.83	2.1-2.4	2.2
Total (Parts A-C)	.87-.91	.89	3.3-3.5	3.3
5. Language Study Skills	.76-.85	.79	2.0-2.4	2.2
6. Arithmetic Computation	.82-.94	.88	2.1-2.7	2.4
7. Arithmetic Problem Solving and Concepts	.90-.95	.92	2.2-2.5	2.4
8. Social Studies Information	.86-.87	.87	3.3-3.5	3.4
9. Social Studies Study Skills	.64-.77	.73	2.2-2.5	2.2
10. Science	.87-.90	.89	2.8-3.3	3.0

Source: Walter N. Durost, *Manual for Interpreting Metropolitan Achievement Tests* (New York: Harcourt, Brace, & World, Inc., 1962), p. 46.

^a Values reported are ranges and medians of four independent estimates of corrected split-half coefficients. Each estimate is based on a random sample ($N = 100$) of grade 6.1 pupils. Each sample was chosen from a single school system, with four school systems being used at each grade level to typify high, low, and average performance on the test.

^b Standard error of measurement in terms of raw score.

Standards for Reliability Coefficients

No arbitrary standards can be established regarding satisfactory levels for reliability coefficients. Obviously, the highest requirements must be set when we are required to make major decisions about *individuals* on the basis of a *single test*; but fortunately such situations are rare. Another situation demanding high test reliability is one where we want to interpret intraindividual differences, that is, *differences between individual scores* on tests that measure various components of scholastic aptitude, musical ability, and the like. The lowest demands on reliability would be made when we are comparing the average scores for large groups. For example, in the problem described in Chapter 1, in which we are comparing the

average spelling achievement of several classes using teaching machines with the average for several classes using standard methods, a less reliable test would be acceptable.

On the basis of certain assumptions²² concerning the accuracy with which a test should discriminate between groups and between individuals, Kelley²³ derived the following minimum reliability coefficients for tests used for different purposes.

PURPOSE FOR WHICH SCORES ARE USED	MINIMUM RELIABILITY COEFFICIENT
To evaluate level of group accomplishment	.50
To evaluate differences in level of group accomplishment in two or more performances	.90
To evaluate level of individual accomplishment	.94
To evaluate differences in level of individual accomplishment in two or more performances	.98

These values have been widely quoted. However, there has been a growing recognition that one can often use to advantage tests with reliability coefficients below these minimums.

In practice, one attempts to find a test that equals or surpasses in reliability the values typically attained in that field of measurement. For example, we should not reject a test of listening comprehension with a reliability coefficient below .94 if our only substitute for such testing is to rely on measures, such as observations and rating, which are far less reliable.

Application of arbitrary standards in the selection of tests can result in unwise decisions. For example, we might select a test with a reliability coefficient of .95 (based on the split-halves method and a combined student population from several grade levels) and reject another test with a reliability coefficient of .92 (obtained by administering alternate forms at a two-weeks interval to students of a single grade level in five different communities and taking the median coefficient). The reliability coefficient for the first test is spuriously high because of the heterogeneity of the group on which the reliability coefficient was computed. The latter test would naturally have a lower reliability coefficient because the *equivalence-stability* coefficient measures all sources of error variance, and the coefficients were computed for several relatively homogeneous groups. The

²² Kelley assumed that a test should permit discrimination of differences in an attribute as small as one-fourth the standard deviation for a grade-level group, with chances of five to one of being correct about the direction of the difference.

²³ T. L. Kelley, *Interpretation of Educational Measurements* (New York: Harcourt, Brace, & World, Inc., 1927).

reliability coefficient of the first test might prove to be considerably lower than .92, if its reliability coefficient were computed under the same exacting conditions as were used for the second test.

IMPROVING THE RELIABILITY OF TEST SCORES

We have already considered some of the factors that affect the size of reliability coefficients—factors that must be taken into account when we are comparing reliability coefficients computed by different methods with groups varying in heterogeneity. We have not yet considered the ways in which the reliability of test scores can be increased.

Increasing Length of Test or Size of Sample

Tests always constitute limited samples of behavior. Hence, a basic approach to improving reliability of scores is to increase the size of the sample. If we check Table A.2, we see that if a test with a reliability coefficient of only .82 is doubled in length, the estimated reliability coefficient of the longer test would be .90; if it were tripled in length, the estimated reliability coefficient would be .93. We can see that reliability increases with size of sample, but that the increase is rather slow. If we are constructing a test, we have to ask (1) whether we can double or triple the number of items and still maintain item quality, and (2) whether the increase in reliability justifies the additional testing time. If we are just interested in seeing how well the group as a whole is doing, a shorter, less reliable test would be adequate.

One application of this principle is a clear-cut one because the quality of the “additional items” would be the same. That is, if we find that one published test has the desired emphasis with respect to subject content but has a reliability coefficient of only .82, we can administer both forms of this test and compute average student scores on the two forms. Thus, we would have a longer test (form A plus form B) that would have a reliability of .90. Using these combined forms to evaluate student achievement in the local educational program would undoubtedly be better than using another test with a reliability coefficient of .90 but with subject content that was not nearly so appropriate to the local pattern of emphasis within a subject field.

Increasing Objectivity in Scoring

Another major factor affecting reliability of measurement is subjectivity of judgment. We tend to get low reliability coefficients for rating scales,

essay tests, ratings of students' products in shop and homemaking, and the like. When a test is objectively scored, this objectivity improves the consistency of measurement.

The objectivity of a measurement depends on the degree to which personal subjective judgment has been minimized in the scoring process. In a multiple-choice test that can be scored by machine, or by a clear-cut scoring key for right-and-wrong answers, the ideal in objectivity is achieved. If some judgment is involved in scoring, but the manual gives fairly precise rules that increase scorer agreement, scoring still remains fairly objective. For example, in the scoring of our 25-word spelling test, scorer agreement would be increased and test scores would be more consistent and reliable if we specified that:

1. A student should not be penalized for failing to dot an *i* or cross a *t*
2. Since spelling, rather than handwriting, is being measured, distinction between *a*'s and *o*'s (and other easily confused letters) should be made on the basis of noting the student's usual ways of writing these letters, as shown in his writing of "personal information" items at the top of his test paper
3. Any clear correction in the spelling of a word should be credited.

These rules, or similar ones, would increase interscorer agreement and make the test scores more objective and reliable. In tests that require judgment in scoring, many examples should be given of responses that should or should not be credited; for example, in the vocabulary test of the Stanford-Binet, types of definitions for each word that would be given "full-credit," "half-credit," or "no credit," have been included.

Maximizing Consistency in Test Administration

All estimates of reliability, except the internal consistency methods, reflect the error variance that is introduced by inconsistencies in test administration, such as deviations in timing, procedure, or student interpretation of test requirements. In other words, a test that has been designed to remove ambiguities for the test administrator and the students will tend to produce more consistent results from test to retest, or from one form to another. For example, we could improve the reliability of our spelling test scores (and decrease the scoring time) by dictating the following advice to students at the end of the test.

Make sure that your score on this spelling test is not lowered by careless handwriting. Look back over your work; dot your *i*'s and cross your *t*'s. Rewrite any word in which the letters are carelessly written so that an *a* might be confused with an *o*, an *m* for an *n*, or an *e* for an *i*.

Selecting Tests of an Appropriate Level of Difficulty for Students

When tests are administered to very heterogeneous groups of students, as is so frequently done in city-wide testing programs, some of the students will find the test so difficult that they will do a great deal of guessing; hence the consistency of these students' scores from test to retest, or from one form to another, will be low.

For very able students or groups, the test selected for city-wide use may be ineffective in differentiating *among* them. Students in able groups get most of the test items right; for each able student, his score might be interpreted as a constant (the number of items of low and average difficulty) plus the individual score he gets on the small proportion of more difficult items. It is as if the test had been shortened to the few difficult items that differentiate among able students, that is, which some able students answer correctly while others do not. If the reliability coefficient of such a test were computed on classes of able students, it would be low, reflecting the brevity of the test for these students and the inconsistencies in their ranks from one "short test" to another.

A desirable approach to this difficult problem is to investigate the consistency of measurement at different score levels. Then the standard errors of measurement for different score levels can be reflected in the norms table, as the Educational Testing Service has done in their use of percentile bands for the STEP tests. However, if we wish to *reduce* the error of measurement, rather than just take it into account in interpretation, it may be best to replace the single city-wide test by two or more tests that are geared to the different levels of achievement of students.

The Educational Testing Service has planned its STEP test series so that STEP tests of two or more levels can be administered to a group at the same time. For example, levels 3 and 4 (designed respectively for grades 7-9 and grades 4-6) can be administered together in a single classroom, with the more able sixth-grade students taking level 3, while other students in the same classroom take level 4 at the same time. The directions and time limits are identical, and the test materials do not specify the grade levels for which they are designed.

Table 3.10 illustrates excellent procedures in computing the standard errors of measurement at different score levels. A large number of students at each grade level were administered both form A and form B of the *Lorge-Thorndike Intelligence Tests*. The average raw score on forms A and B was computed for each student. Then subgroups were formed of students whose average scores fell at each raw-score level (for example, those with average raw scores of approximately 15, 20, and the like). Then for the students in each raw-score subgroup, differences between form A and form B scores were tallied and the *SD*'s computed. These

SD's are the standard errors of measurement for each score level. If we examine the fifth column in Table 3.10, (for 3d grade pupils, verbal battery), we see that the standard error is approximately twice as large for pupils obtaining raw scores of 20 and below, than for those scoring in the middle range. More reliable estimates of performance on this type of intelligence-test items would probably be obtained by administering the next lower level of the test to these low-scoring pupils.

Table 3.10

Standard Error of Measurement of the Lorge-Thorndike Intelligence Tests
(Grades 3-5) at Selected Raw Score Levels

AVERAGE RAW SCORE	STANDARD ERRORS OF MEASUREMENT (IN IQ POINTS)						
	GRADES:	Nonverbal Battery			Verbal Battery		
		3	4	5	3	4	5
15		8.7	8.1	6.5	6.6		
20		7.9	7.6	5.9	5.8	6.0	6.0
25		7.0	7.2	5.4	5.1	5.4	5.6
30		6.2	6.9	5.2	4.5	4.9	5.2
35		5.8	6.6	5.6	3.9	4.5	5.0
40		5.6	6.5	6.2	3.5	4.2	5.0
45		5.5	6.7	6.6	3.2	4.1	5.1
50		5.7	7.0	6.7	3.0	4.2	5.2
55		6.0	7.4	6.3	3.5	4.4	5.1
60		6.3	7.8	5.8	4.3	4.5	4.8
65		6.9	8.2		5.0	4.6	4.5
70		7.8	8.6		5.5	4.7	4.0
75					5.9		
Weighted average standard error		6.2	7.1	6.1	4.4	4.6	5.1
Reliability coefficients (equivalent forms method)		.85	.80	.85	.92	.92	.90
Number of cases		2659	1419	834	2659	1419	834

Source: Adapted by permission of the publisher from Irving Lorge and Robert L. Thorndike, *Technical Manual*, rev. ed. (Boston: Houghton Mifflin Company, 1962), p. 11.

SUMMARY STATEMENT

A person's test score summarizes data on his performance on a *sampling* of tasks or test items. The concept of reliability is concerned with the consistency of measurement, or the extent to which an individual's scores vary from one sample to another of the same type of behavior.

If the same test is readministered on two different occasions, we obtain data on variance in scores due to temporal variations in the examinees. If two forms of a test are administered on the same occasion, we obtain data on variance in scores due to specificity of the samplings of test items.

The various sources of inconsistency in examinee behavior from one testing to another are summarized in Table 3.1; while Table 3.4 clarifies the extent to which each of these sources of variance is taken into account by the different approaches used in the estimation of test reliability.

Tables for approximating Kuder-Richardson reliability coefficients and standard errors were presented in order to enable students to utilize and interpret these measures without necessarily developing proficiency in computation. The many factors involved in making comparisons between reliability coefficients presented in test manuals were considered, namely (1) the method used in estimating reliability and (2) the ability range of the groups studied.

Although reliability coefficients are most useful in assessing the comparative reliability of different tests, the standard error is more valuable in the interpretation of test scores for individuals.

The reliability of differences between pairs of scores is much less than the reliability of either score. If two tests measure closely related abilities, their reliability coefficients must meet high standards if one is to interpret difference scores with a reasonable degree of confidence. For examples of this relationship, the reader is referred to Table 3.8.

The reliability of a test, or the consistency of student scores from one test sample to another, depend largely on the length of the test, the homogeneity of the universe sampled, and the objectivity of test scoring.

SELECTED REFERENCES

- CRONBACH, LEE J., AND GOLDINE C. GLESER, *Psychological Tests and Personnel Decisions*. Urbana, Ill.: University of Illinois Press, 1957.
- DIEDERICH, PAUL B., *Short-cut Statistics for Teacher-Made Tests*. Evaluation and Advisory Service Series No. 5. Princeton, N.J.: Educational Testing Service, 1960. Available on request.
- LORD, FREDERIC M., "Tests of the Same Length Do Have the Same Standard Error of Measurement," *Educational and Psychological Measurement*, vol. 19 (Summer 1959), pp. 233-239.
- , "The Utilization of Unreliable Difference Scores," *Journal of Educational Psychology*, vol. 49 (June 1958), pp. 150-152.
- SUPER, DONALD E., AND JOHN O. CRITES, *Appraising Vocational Fitness*, rev. ed. New York: Harper & Row, Publishers, Inc., 1962, Chapter 3.
- THORNDIKE, ROBERT L., "Reliability," in E. F. Lindquist, ed., *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, pp. 560-620.
- WESMAN, ALEXANDER G., "Better Than Chance," *Test Service Bulletin* No. 45. New York: The Psychological Corporation, 1953. Available on request.
- , "Reliability and Confidence," *Test Service Bulletin* No. 44. New York: The Psychological Corporation, 1952. Available on request.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. What are the major factors that influence the reliability of a test?
2. When is a test objective? How is objectivity related to reliability?
3. Discuss the relative merits of each of three methods used in determining the reliability of a test.
4. Study the manuals for two or more standardized tests, noting the reliability for all subtests. Summarize and evaluate the evidence presented in terms of using the test to measure the achievement of groups. Of individuals. How reliable are the differences between scores on various pairs of subtests?
5. Illustrate how a teacher should use the standard error of measurement in his interpretation of test scores.
6. Three standardized tests have reliability coefficients of .70, .80, and .90 respectively. How many forms of each test would need to be used to yield sufficiently reliable results for individual measurement? How did you reach your conclusion?
7. Discuss the importance of selecting or developing a test of suitable difficulty level for a specific group of students.
8. Why is the reliability coefficient obtained by the "odd-even" method considered to be a "coefficient of equivalence," even though only one form of the test is used?

The term "validity" is used to apply to a test's value as a basis for making judgments about examinees. A single test may be used for making several types of judgments; its validity may be high for one purpose, moderate for another, and low for still another. Hence, we cannot speak of a test as having high or low validity without specifying the purpose for which it is to be used.

The term "purpose" is best interpreted as including both the type of judgment to be made and the nature of the group involved. A test in business English may be valid for differentiating among high school students to make judgments basic to grading. The same test may make little contribution to the goal of differentiating among applicants with respect to predicted success in secretarial positions. Hence, validity is always validity for a specific purpose (to aid in making a specific type of judgment concerning members of a specific group).

Validity has two major aspects—reliability and relevance. For a test to be valid, that is, to provide a sound basis for judgments, it must measure "something" with reasonable reliability, and that "something" must either be a sample of the behavior we wish to measure or it must have demonstrated relevance to that behavior. Reliability, or the consistency of measurement, was studied in Chapter 3. In this chapter, we will be chiefly concerned with relevance—the relationship of scores on the test to the criterion behavior in which we are really interested.

TESTS AS DIRECT OR INDIRECT MEASURES OF CRITERION BEHAVIOR

Tests as Direct, Unbiased Samplings of Criterion Behavior

When the content of a test is a random sampling of a defined area of content, relevance is not a problem. An example would be the spelling

test designed for use in problem 3 on the use of teaching machines in spelling instruction. The criterion behavior we wish to measure in this situation is student performance on the population of 500 words. Since the sample is a random one, it is unbiased and perfectly relevant. Hence the validity of this test is determined entirely by its reliability (which depends chiefly on size of sample and objectivity of scoring). It is rare, however, that test content constitutes a random sampling of criterion behavior. Whenever the sampling is not random, human judgment enters into the selection of learnings to be tested.

Tests as Indirect Measures of Criterion Behavior

In most tests, we do not sample the criterion behavior in which we are really interested. For purposes of efficiency in scoring, we introduce multiple-choice items when we are really interested in how well students can compute, punctuate, spell, and the like. Hence, we must make statistical studies to determine how much irrelevant variance has been introduced by our use of such indirect methods. We can correlate scores on our indirect measure with scores on the same test items, presented in the direct manner, in which students actually work out the arithmetic problems, punctuate the sentences, or spell the words.¹ In this situation our criterion is the score on the test that demands student recall of information, the actual working of problems, or spelling of words.²

In many situations it is very difficult to assess the validity of a test as

¹ Hopkins uses the term "coincident validity" or "extrinsic reliability" for the correlation between scores on a multiple-choice test and scores on the same test, or an equivalent form of the same test, administered as supply-type items (that is, excluding the alternative responses for items). Kenneth D. Hopkins, "Validity Concomitants of Various Scoring Procedures Which Attenuate the Effects of Response Sets and Chance," unpublished doctoral thesis, University of Southern California, 1961.

² This type of correlation coefficient may be perceived as a validity coefficient in that it takes into account stable but invalid variance associated with the use of selection-type items. Such variance lowers the relevance of the test scores to the criterion behavior, in which we are really interested. Such coefficients may, however, be perceived as reliability coefficients in that they reflect types of error variance. Reliability coefficients are estimates of the consistency of measurement by *methods that are maximally similar*, while validity is concerned with agreement or convergence among scores that are obtained by quite different methods. On such a continuum of similarity vs. diversity of methods, the type of coefficient we are considering would, in the opinion of the author, lie farther from the maximal similarity-of-method end of the continuum than the subdivided-test method, or even the alternate-forms method, but would still be more of a validity coefficient than a reliability coefficient. This is a subjective judgment, however, and Hopkins (footnote 1) prefers the term "extrinsic reliability." This explanation is intended more to illustrate the fine gradations between methods of studying error variance than as support for the choice of one term or the other.

an estimate of ultimate criterion behavior. For example, we may wish to make judgments about a person's driving ability on the basis of his performance on the test he takes to obtain his driver's license. The validation problem becomes one of relating driving test scores or ratings to the ultimate criterion of "success in driving." However, we face a difficult problem in that it is almost impossible to obtain reliable measures of the ultimate criterion, that is, scores that represent an objective evaluation of day-by-day criterion behavior as a driver.

Tests as Predictors of Future Criterion Behavior

Let us consider another practical problem, that of making predictions about students' *future* success in an activity, for example, clerical work. If we want to evaluate a test as a basis for making such judgments, actual success in clerical jobs would be the "ultimate criterion." We may decide, however, to use teachers' marks in a clerical practice course as an "intermediate criterion" for judging the validity of a clerical aptitude test. If so, we should study the relevance and reliability of this intermediate criterion. Teachers' marks in the clerical practice course may have low relevance to the ultimate criterion of success on the job. In studying relevance, we would investigate the extent to which the clerical tasks and standards of performance in the course are representative of those on the job. Both relevance and reliability would be affected by the degree to which teachers based their marks on subjective general impressions and unconscious bias, rather than on objective data on student performance.

The test-maker frequently has to check his test against an "intermediate criterion" that may not be closely related to the "ultimate criterion" in which he is interested. For example, success in a clerical practice course may have only a low or moderate relationship with the ultimate criterion of "success on the job." Obtaining suitable measures of criterion behavior, to use as standards for validating tests, has been one of the most difficult problems in test validation; it has stimulated study and discussion among outstanding leaders in the field of measurement.³

³ For further treatment of this problem, see Robert Hoppock, ed., "Criteria of Vocational Success—A symposium," *Occupations*, vol. 14 (June 1936), pp. 917-975; R. M. Bellows, "Procedures for Evaluating Vocational Criteria," *Journal of Applied Psychology*, vol. 25 (October 1941), pp. 499-513; Edward E. Cureton, "Validity," in E. F. Lindquist, ed., *Educational Measurement* (Washington, D.C.: American Council on Education, 1951), pp. 621-694; Edwin E. Ghiselli and C. W. Brown, "Analysis of Jobs," *Personnel and Industrial Psychology* (New York: McGraw-Hill Book Company, Inc., 1955), pp. 17-58; D. B. Stuit, "The Effect of the Nature of the Criterion upon the Validity of Aptitude Tests," *Educational and Psychological Measurement*, vol. 7 (Winter 1947), pp. 671-676; Donald E. Super and John O. Crites, *Appraising Vocational Fitness* (New York: Harper & Row, Publishers, Inc., 1962), pp. 32-41; Robert L. Thorndike, *Personnel Selection: Test and Measurement*

Fortunately for the test-maker, some intermediate criteria have significance in their own right. These criteria, in a sense, carry their own labels of success or failure, such as graduation from high school, retention of a job, making at least a C average during the freshman year in college. That is, the ultimate criterion behavior, involving "success on the job," can be exhibited *only* by persons who pass certain intermediate hurdles. A person will have no opportunity to show his performance on the ultimate criterion of success in a specific profession unless he first scores sufficiently high on college admission tests, earns a certain grade-point-average in college and professional school, and passes some type of a licensing examination. Hence student performance on any of these intermediate criteria can serve as a partial basis for validating aptitude tests. Moreover, teachers' marks and supervisors' ratings, even though biased, are socially significant in their effects; and hence the relationship of test scores to these criteria is worthy of study.

We must not assume, however, that the relationship of intermediate criteria to ultimate criteria is unimportant. It is to the advantage of both society and the individuals concerned if we guide into training programs those persons who will *ultimately* be successful in the actual jobs or professions. Success on preliminary training hurdles is necessary but not sufficient.

TYPES OF JUDGMENTS MADE ON THE BASIS OF TEST RESULTS

When we construct or select a test, our chief concern is that the test scores enable us to improve our bases for judgments about the examinees. In the *Technical Recommendations*, four types of judgments that test users desire to make are listed as a basis for clarifying the different types of validity studies which need to be made. These types of judgments, or purposes⁴ in testing, are stated as follows:

1. The test user wishes to *determine how an individual would perform at present* in a given universe of situations of which the test situation constitutes a sample.

Techniques (New York: John Wiley and Sons, Inc., 1949), Chapter 5; H. A. Toops, "The Criterion," *Educational and Psychological Measurement*, vol. 4 (Winter 1944), pp. 271-297.

⁴ In this statement of generalized purposes, only the type of judgment is indicated and no reference is made to the group of individuals, about which such judgments will be made. In evaluating tests for local use, however, both the type of judgment to be made and the nature of the group tested must be considered.

2. The test user wishes to *estimate an individual's present status on some variable external to the test*.
3. The test user wishes to *predict an individual's future performance* (on the test or on some external variable).
4. The test user wishes to infer *the degree to which the individual possesses some trait or quality (construct)*, presumed to be reflected in the test performance.⁵ [Italics added. Items 2 and 3 have been interchanged to correspond with order of presentation in the text and Tables 4.2 through 4.9.]

Each of these four types of judgments involves a different focus of concern with respect to test validity. When we are making judgments of the first type, we are chiefly concerned with *content validity* (how well our test sample represents the universe of criterion behavior); in the second we are interested in *concurrent validity* (how closely test scores are correlated with present criterion behavior); in the third we are concerned with *predictive validity* (how well test scores predict future criterion behavior); for the fourth purpose we are concerned with *construct validity* (how well our test seems to measure the hypothesized trait—as shown by the effectiveness of the test in differentiating among groups that are presumed to differ with respect to the trait, and also by the relationship of test scores to predicted behavior in natural or specially designed situations). Each of these types of validity will be discussed and illustrated.

It is impossible for the authors of a standardized test to provide completely adequate data on the validity of their test for all purposes (judgments and groups) for which test users might conceivably employ the test. The authors should, however, provide data that enable the test user to judge whether the test is *likely to be valid for a specific purpose*. Once he has selected a test on the basis of such a hypothesis, he should collect local validation data to check on the test's validity as a basis for making the type of judgments he wishes to make about the students he wants to select, classify, or counsel.

CONTENT VALIDITY

Random Sampling of a Universe as a Basis for the First Type of Judgment

The reader will recognize that when we used the local spelling test we wished to make the first type of judgment (listed above) in order to assist us in "evaluation of treatments," that is, evaluation of the efficacy of a

⁵ "Technical Recommendations for Psychological Tests and Diagnostic Techniques," Supplement to the *Psychological Bulletin*, vol. 51 (March 1954), p. 213.

specific teaching-machine program. Our test is a random sampling of the spelling words studied; hence our criterion is student performance on the defined universe of 500 words. It is not feasible to use student performance on a 500-word dictation spelling test as our criterion, but we can estimate the validity of the shorter test from its reliability coefficient.

The shorter test is a perfectly relevant sample of criterion behavior except for the sampling errors involved in using only 25 of the 500 words. If the reliability coefficient⁶ is .64, we can estimate that the validity coefficient is .80 (the square root of the reliability coefficient).⁷ This coefficient is an estimate of the correlation between students' scores on one sampling of 25 words and their "true scores" (theoretical scores on the entire universe of words). In other words, when we can exactly define the universe of criterion behaviors we wish to measure, and can obtain unbiased measures of student performance on a sample of this universe, the validity coefficient of the sample test depends entirely upon its reliability.

Achieving Representativeness of Sampling When Random Sampling Is Not Feasible

When we measured student achievement in history in problem 4, we were also concerned with making the first type of judgment; that is, we wished to make inferences, from student performance on sample tests, concerning their knowledge of the history of their state and nation. Defining the universe of knowledge of history, however, is intrinsically more difficult than for spelling. In fact, professional judgment concerning the relative emphasis on different areas is required. Since all the teachers used the same textbook in state history, and the course of study indicated roughly the division of time to be allotted each major area, the teachers could agree on the percentage distributions listed at the bottom, and in the left-hand column, of Table 4.1. On the basis of this sampling plan, they were able to devise a test that fitted the teachers' specifications quite closely.

One can see by comparing the "total" column with the percentages at the left that the specifications for the distribution of items by chronological period were precisely met. With respect to aspects of history, the specification to have 40 percent of the items on the political and military aspects was satisfied; but the plan to have 30 percent for each of the other aspects

⁶ The coefficient of equivalence and stability would be the best method of measuring reliability for this purpose; all sources of error variance are included (Table 3.4).

⁷ It may seem strange that a sample test could correlate higher with the criterion than with another sample test. In this situation, however, the criterion is student performance on the entire universe of words; hence sampling errors are not involved in the criterion measure. The validity coefficient, therefore, is not reduced as much by sampling error, as is the reliability coefficient. (The latter is an r between two sample tests, which is reduced by the sampling error in *both* tests).

Table 4.1
Table of Specifications (or Sampling Plan) for City-Wide Test in
State History (100 Items)

CHRONOLOGICAL PERIOD (AND DESIRED EMPHASIS ON EACH PERIOD)	Number of Items on Each Aspect of History			TOTAL
	POLITICAL AND MILITARY ASPECTS	SOCIAL AND CULTURAL ASPECTS	ECO-NOMIC ASPECTS	
Exploration and colonization (10%)	5	2	3	10
Establishment of state government (15%)	11	2	2	15
Development of the new state (15%)	6	4	5	15
Involvement in the War between the States and the Reconstruction period (10%)	4	2	4	10
Industrial and cultural development (1875-1910) (15%)	3	5	7	15
World War I, the postwar period, and the depression years (15%)	4	4	7	15
World War II to the present (20%)	7	6	7	20
Total number of items (by aspects of history)	40	25	35	100
Desired emphasis on each aspect	40%	30%	30%	

was only approximately achieved. Note that no attempt was made to balance the questions on aspects of history for *each* chronological period.

Appraising the Content Validity of Standardized Tests for Local Use

With the test of United States history, teachers were also concerned about the first type of validity, content validity, that is, the extent to which the content of different standardized tests *represented* the universe of content they wished to sample. In this case, they decided on a two-way table of specifications *involving both type of objective and area of subject matter*.⁸ They decided that only 40-50 percent of the items should be concerned with memory of *knowledge*; that 30-40 percent should require that the student demonstrate his *comprehension* of what he had learned

⁸ For an example of such a table of specifications (involving both objectives and content) the reader is referred to Table 10.1 of Chapter 10.

through his ability to interpret trends, explain cause-effect relationships, and the like; and that 10–30 percent of the items should involve such higher abilities as *application*, *analysis*, and *evaluation*. The terms in italics refer to five of the six categories in the taxonomy of objectives,⁹ which will be reviewed in Chapter 11.

On the basis of the pooled judgments of teachers concerning their desired emphases, the committee developed guide lines for examining a number of standardized history tests for their content validity for their purposes. The committee discovered that some test manuals gave adequate information in their own tables of specifications so that it was comparatively easy to judge the relevance of the test to local curricular emphases. For other tests it was necessary to go through the test, item by item, attempting to classify the items according to objective and content.

The committee concluded that no test actually fit their local specifications closely. However, two tests were worthy of further study in that they had adequate reliability coefficients, fair content validity, and met other practical criteria regarding ease of administration, scoring, and the like. One of these tests corresponded quite well with the teachers' proposed emphasis with respect to content but not with respect to objectives. This test included 70 percent items of the "knowledge" type, 30 percent "comprehension," and none representing the other types of objectives. The second test, although it included far too little content on the history of the United States during the twentieth century, had a better distribution with respect to the objectives measured.

The committee decided that, since the community was most interested in knowledge, they would choose the first test, which approximated their desired distribution with respect to content. Then they would devise a supplementary test of their own that would include items which tapped the higher abilities of interpretation, application, and so forth. They realized that the second test, with its minimal emphasis on recent history, would not constitute a fair basis for making judgments about the effectiveness of their instruction in furthering the knowledge objectives of American history.

Summary of Procedures Involved in Constructing and Validating Tests for the First Purpose¹⁰

Table 4.2 is concerned with the construction and validation of tests designed to serve the first purpose. This table not only gives further examples of content validity, but specifies the procedures that should be

⁹ Benjamin S. Bloom, ed., *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain* (New York: David McKay Company, Inc., 1956).

¹⁰ As listed in the "Technical Recommendations for Psychological Tests and Diagnostic Tests," *op. cit.*, and quoted on pages 106–107 of this textbook.

Table 4.2
Construction and Validation of Tests to Be Used To Serve the First Purpose^a—Sampling Present Performance

AIM IN TESTING (TYPE OF JUDGMENT TO BE MADE)

To estimate how a person would perform, at present, in a defined "universe" of situations, of which the test constitutes a sample
EXAMPLES: To estimate how a person drives a car in typical urban traffic conditions; to estimate how accurately a child spells the words he has studied this school year; to estimate how many of the words usually included in first-grade readers a child can recognize.

TYPES OF TESTS USED FOR THIS PURPOSE

Standardized tests of achievement, teacher-made achievement tests, work-samples (such as a sample of driving performance), and the like. All these tests provide a record of individual performance on a sampling of a defined universe of situations.

GENERALIZED PROCEDURES FOR THE DEVELOPMENT OF TESTS FOR THIS PURPOSE

1. Define the universe of content and situations to be sampled.
 - a. Restrict the universe to be sampled in terms of
 - (1) Types of inferences you wish to make from test scores^b
 - (2) Feasibility of obtaining data^c
 - b. Restrict the universe to be sampled in terms of aspects of behavior to be studied.^d
2. Sample the universe by procedures that can be clearly described.
 - a. If one is sampling a finite, clearly defined universe of items, random sampling can be used.^e
 - b. In the construction of most achievement tests, the universe of content and abilities to be sampled cannot be as clearly defined. Hence, a basis for selecting a stratified sampling of knowledges, skills, and the like should be decided upon. The teacher, or a professional group, should decide on the relative emphasis to be given each area of content and each major objective.
 - (1) For a teacher-made test, the teacher himself can make these judgments although he may find it desirable to check his judgments with those of coworkers in the same subject area.
 - (2) For a standardized test, pooling the judgment of professionally trained teachers and supervisors in the subject field^f is highly advisable.
 - c. After items have been written and tried out with students, the test constructor should select items so as to maintain the approximate distribution of items (by area and objective) as was planned in advance when the test "blueprint" was designed.

Table 4.2 (Continued)
Construction and Validation of Tests to Be Used To Serve the First Purpose^a—Sampling Present Performance

3. Unless the total test is highly homogeneous in content, the items should be grouped into relatively homogeneous subtests.
 - a. Since many achievement test batteries are intended to sample a very broad universe of learnings, the items of the battery may measure a number of abilities that are not highly interrelated. For meaningful interpretation of test scores, therefore, test items should be grouped into subtests that sample relatively homogeneous components of the universe sampled.⁵
 - b. For any subtest (group of items for which a score is computed), evidence should be presented about the internal consistency of that subtest; that is, evidence concerning the reliability of subtest scores is essential if we are to avoid unwarranted inferences about the relative achievement of individual students in different areas.^h

INFORMATION THAT SHOULD BE PROVIDED IN A TEST MANUAL REGARDING THE CONTENT VALIDITY OF A TEST

The sources and criteria used to ensure that the test sample is representative of performance in the defined universe should be clearly stated. The manual should indicate:

1. The sources from which items were drawn.¹
2. The criteria for inclusion or exclusion of items.

The blueprint or table of specifications for an achievement test (or other sampling of performance) should contain sufficient information regarding

 - a. The bases for selecting test items
 - b. The number of items for each content area or skill.
3. Information should also be presented concerning procedures used to increase the efficiency of measurement so that optimum use could be made of available testing time. It is efficient to eliminate from the preliminary edition of a standardized achievement test (or semester examination, or any other test used to rank students):
 - a. Items that are too easy or too difficult for the group for which the test was designed.
 - b. Items that make little or no contribution to differentiating between high-achieving and low-achieving students on the variable sampled.
 - c. Items that are trivial or irrelevant.

Professional judgment, based on data from a try-out of the test, should be used in making such decisions.

Table 4.2 (Continued)
Construction and Validation of Tests to Be Used To Serve the First Purpose^a—Sampling Present Performance

^a The four aims, listed on pp. 106–107, and used as the basis of organization of Tables 4.2, 4.3, 4.5, and 4.9, are rephrasings of the four aims of testing, listed in "Technical Recommendations for Psychological Tests and Diagnostic Techniques," Supplement to the *Psychological Bulletin*, vol. 51 (March 1954), p. 213.

^b For example, if we wish to make judgments about a student's learning of assigned spelling words at his grade level, rather than his over-all level of spelling ability, we will sample only the words in the speller for his grade.

^c For example, we may not find it feasible to include mountain-driving behavior in the test sample for applicants for driving licenses. Hence, we restrict the universe sampled, and recognize that we will not be justified in making inferences about unmeasured aspects of driving ability.

^d For example, if we are studying a sample of first-graders' performance in reading, we may decide to note only the number of words recognized; we will not record information about whether or not the pupil reads with expression, whether he makes frequent pauses and retrogressions, or whether he seems tense.

^e For example, in devising the local spelling test, mentioned earlier, the teachers selected every 20th word from a universe of 500 words.

^f The Educational Testing Service has recognized that the measurement expert should not make decisions concerning the proportional emphasis to be given to different objectives and different content areas. For

each subject area of the STEP tests, national professional organizations of teachers were consulted for their recommendations concerning personnel to work on committees that would draft the blueprint or table of specifications for each test. In fact, much of the preliminary item writing was done by members of these representative committees. The California Test Bureau and many other publishers have also obtained the reactions of teachers and supervisors concerning test items to be included in achievement tests.

^g For example, if we include items on handwriting, spelling, mechanics of English, and language usage in one omnibus test of language, we cannot make as meaningful inferences about student achievement as we could if scores were available on fairly homogeneous subtests.

^h Criterion C4.3 of the Technical Recommendations reads as follows: "If items are regarded as a sample from a universe, a coefficient of internal consistency should be reported for each descriptive score, to demonstrate the extent to which the score is saturated with common factors. ESSENTIAL." Quoted from "Technical Recommendations for Psychological Tests and Diagnostic Techniques," Supplement to the *Psychological Bulletin*, vol. 51 (March 1954), p. 20.

ⁱ For example, the person using the test should know what spellers (or standard spelling lists) were sampled for a spelling test; what vocabulary lists were used as the basis for selecting words for a measure of students' vocabulary level, and the like. If a standardized test has based its items on the content of certain currently used textbooks, these books should be listed in the manual, with their dates of publication.

followed in developing tests which sample a universe of knowledge or skills. In the last section of the table are listed the types of information that should be given in the manual of a published test so that the user could judge the content validity of the test for his own purposes.

CONCURRENT VALIDITY

The most typical example of the second purpose in testing (estimating the individual's status on some attribute external to the test) arises when a test is being used as a more economical, convenient substitute for accepted appraisal procedures. A group intelligence test may be substituted for an individual test; or a multiple-choice spelling test may be substituted for the students' actual spelling of words.

When a shortcut substitute for some more elaborate standard method of measurement is proposed, the question of the validity of the substitute method does arise with logical legitimacy. In such a situation the concept of validity is simple, and the meaning of the term is clear.¹¹

The Construction of a Test Specially Designed for the Second Purpose

An excellent example of the use of concurrent validity studies in test construction is the work done by the College Entrance Examination Board in developing objective tests of composition skills.¹² The Board had long been concerned about the time involved in judging essays and the subjectivity of judges' ratings. Hence, their staff has engaged in considerable research to develop objective tests, designed to correlate highly with the criterion of student performance in essay writing.

As a basis for obtaining criterion scores, each student in the research group wrote five different essays, each of which was given ratings by five judges. The criterion score was a composite of 25 ratings (five judgments on each of five essays). The investigators made sure that each judge was uninformed about the ratings assigned essays by other judges, and uninformed about students' scores on the objective tests of composition skills. These steps were taken to avoid criterion contamination (that is, to avoid criterion scores being affected by the rater's knowledge of test scores or other relevant data that might affect ratings). The reliability coefficient for the criterion scores was .84.

¹¹ Robert L. Ebel, "Must All Tests Be Valid?" *The American Psychologist*, vol. 16 (October 1961), pp. 641-642.

¹² *Annual Report, 1961-62* (Princeton, N. J.: Educational Testing Service, 1962), p. 99.

Several types of items were found to have concurrent validity and were included in the tests because of their correlation with the composite criterion score on essay writing, for example:

1. Multiple-choice questions requiring the student to choose, from a number of alternatives, the best expression for an indicated word or phrase
2. Multiple-choice questions requiring the student to classify sentences according to whether they (a) contained an error in diction, (b) were verbose or redundant, (c) contained clichés or abused metaphors, or (d) contained faulty grammar
3. Exercises requiring the student to select the appropriate line to complete a poem and indicate whether each of the other alternatives is (a) inappropriate in rhythm or meter, (b) inappropriate in style or tone, or (c) inappropriate in meaning
4. Prose exercises similar to (3) above, except that the rejected alternatives are to be classified as (a) inappropriate in meaning, (b) inappropriate in tone or diction, or (c) grammatically defective.¹³

Summary of Procedures Involved in Constructing and Validating Tests for the Second Purpose

In developing this test of English composition skills, the authors' purpose and procedures paralleled those listed in Table 4.3 on concurrent validity. This was not a situation in which test authors could sample a defined universe of content and abilities. Instead, they had to (1) devise an adequate criterion measure and obtain criterion scores for a large number of students, (2) devise a large number of items on the basis of hypotheses about test items likely to tap the abilities used in writing essays, (3) study examinee performance on each item in relation to criterion data, and (4) select items for the revised test that maximized the relationship between test scores and the criterion data.

Concurrent Validity Data as a Partial Substitute for Data on Predictive Validity

Since concurrent validity coefficients, in which we relate test scores to *present* performance, can be obtained with less expense and delay than predictive validity coefficients (involving *future* performance), many aptitude test manuals present concurrent validity data only. Evidence that a test has fairly high concurrent validity does *not* justify the assumption that the test also has predictive validity. In fact, Maurer¹⁴ discovered that

¹³ Adapted from *A Description of the College Board Achievement Tests* (Princeton, N. J.: Educational Testing Service, 1962), pp. 19–39.

¹⁴ Katherine M. Maurer, *Intellectual Status at Maturity as a Criterion for Selecting Items in Preschool Tests* (Minneapolis, Minn.: University of Minnesota Press, 1946).

Table 4.3
Construction and Validation of Tests Used to Serve the Second Purpose—Indirect Assessment of Criterion Data

AIM IN TESTING (TYPE OF JUDGMENT TO BE MADE)

To estimate a person's present status with respect to some attribute external to the test; usually to obtain more easily, quickly, and inexpensively estimates of the examinee's present status with respect to some attribute that cannot feasibly be measured by a more direct method.

EXAMPLES: To estimate a person's mental age on an individual test (for example, the *Wechsler Intelligence Scale for Children*) by using a short form of that test, or a group test; to estimate a college applicant's ability in essay writing by a series of objective and semiobjective exercises presumed to test similar abilities; to estimate a student's ability to plan experiments by asking him to rank suggested experiments in their order of value for testing certain hypotheses.

TYPES OF TESTS USED FOR THIS PURPOSE

Short-form individual intelligence tests; group intelligence tests; apparatus tests used to estimate skills, such as driving skills, under simulated conditions; multiple-choice tests of such skills as spelling, ability to solve chemical equations, and the like; projective tests or objectively-scored personality inventories, used as a substitute for individual interviews by psychologists (in the tentative assignment of patients to diagnostic categories, or in the selection of students who should be referred to a counselor or psychologist).^a

GENERALIZED PROCEDURES FOR THE DEVELOPMENT OF TESTS FOR THIS PURPOSE

1. Obtain the direct or criterion measure of the attribute for a group of students, employees, patients, or other subjects.^b
2. Devise a large number of test items on the basis of hypotheses about test items likely to tap the abilities or traits represented in the criterion.^c
3. Administer this preliminary edition of the test to persons for whom criterion data are, or will be, available.
4. Study examinee performance on each item of the preliminary test in relation to criterion data. This study could be made by
 - a. Computing the correlation between criterion scores and performance on each item.^d
 - b. Comparing the performance on items for criterion groups, for example, diagnosed neurotics v.s. a random sample of people in general.
5. Select items for the revised test that maximize the relationship between scores on the test and the criterion data.^e

INFORMATION THAT SHOULD BE PROVIDED IN A TEST MANUAL REGARDING THE CONCURRENT VALIDITY OF A TEST

1. Data concerning the adequacy of the criterion should be given in the test manual.^f

Table 4.3 (Continued)
Construction and Validation of Tests Used to Serve the Second Purpose—Indirect Assessment of Criterion Data

2. Validity coefficients (correlations between test and criterion scores) should be presented for groups that are similar to those on which this test is likely to be used. These coefficients should be computed on a different group or groups of examinees than the one used in item selection. The validity coefficients for the original group would be spuriously high because we would be testing the validity of the instrument on the same group for which we did the item selection.
3. The test manual should present data that help the user to judge the reliability of the estimates he makes of criterion data from test scores.^a
4. Unless the concurrent validity of the test is approximately the same throughout the full range of scores, the standard error of estimate should be given for different score levels.^b
5. Since some criteria, such as ratings, are very unreliable, it is fairly common practice to report concurrent validity coefficients that have been corrected for the unreliability of the criterion scores. Such a corrected coefficient measures the *theoretical* relationship between test scores and perfectly reliable criterion scores.^c If a corrected coefficient is given, the uncorrected one should also be given so as not to give a misleading impression of the effectiveness with which a short-cut measure predicts actual criterion scores. Validity coefficients should not be corrected for the unreliability of the short-cut or substitute test.

^a Ordinarily, such measures are used only as preliminary screening devices to select those persons most likely to need the interviewing time of a highly paid psychologist or psychiatrist.

^b For example, the individually administered WISC can be given to a number of subjects to obtain criterion scores. Or students can be asked to write two or more essays, each of which will be rated by several judges, making their judgments independently; the composite score based on these independent judgments would constitute a criterion score on essay writing.

^c For example, the author of a test on neuroticism would construct items that, according to his hypotheses, are likely to be answered differently by the neurotic person (as compared with people in general). The author of an objective test of composition skills might devise items on: correctness of usage, ability to find and correct errors in

first-draft material, ability to rearrange sentences in scrambled paragraphs, and the like.

^d Special techniques are available for minimizing the amount of work required in computing correlation coefficients when one or both of the variables are dichotomous (for example, a student's response to an item is either "right" or "wrong," scored either +1 or 0). See any standard textbook in statistics, for example, Quinn McNemar, *Psychological Statistics*, 3d ed. (New York: John Wiley and Sons, Inc., 1962), Chapter 12.

^e For example, let us assume that we are developing a short form of the WISC for use with brain-damaged children of elementary school age. If we wanted to use this short form to get a rapid assessment of the child's mental age, we would select items that showed the highest correlation with score on the total test. If our purpose were

Table 4.3 (Continued)
Construction and Validation of Tests Used to Serve the Second Purpose—Indirect Assessment of Criterion Data

to develop a test that would aid in diagnosing brain damage, we would select items which gave us maximum differentiation between brain-damaged and normal children. In one case, our criterion would be score on the total WISC; in the other case, our criterion would be membership in the "brain-damaged" or normal group, on the basis of classification by a neurologist.

^c Although the adequacy of the WISC as a criterion of mental ability has been well established, such a criterion as a "composite grade on essays" has not. Hence, in the latter case, information should be given in the test manual concerning selection of topics for essays, training of scorers, consistency of grading between scorers and the like. When ratings are used as criteria, it is especially important that rater reliability and rater bias be studied; for example, some raters may be grading essays chiefly in terms of mechanics; others may be biased by neatness; still others may be emphasizing originality. Directions to raters, and training in their work, help to reduce personal bias.

^e For example, if the user is estimating students' IQ's on the total test from their scores on a shorter version, he should be given information concerning the standard error of estimate, so that he can know the "confidence limits" within which criterion scores can be expected to vary. If criterion groups have been compared (such as neurotic and normal, or brain-damaged and normal), information should be given concerning the likelihood of misclassification if test data are used as a short-cut substitute for criterion data.

^h For example, a brief form of a mental test might measure with insufficient reliability at either the low or high extreme of the score range; if so, the brief form would not constitute a valid basis for estimating full-scale scores for such low-scoring or high-scoring examinees.

ⁱ The formula for correcting a correlation coefficient for the unreliability of one or both variables is called the "correction for attenuation." The formula for this correction is given in any standard statistics textbook; it may be found in McNemar, *op. cit.*, p. 153.

certain items in intelligence tests for young children had concurrent validity (in the sense of being related to other evidences of ability at that time); while different items proved to have greater predictive validity (in that they were associated with high future performance).

Presentation of concurrent validity data for aptitude tests does not relieve the authors and publishers of their responsibility to follow up students, as they go into college and into vocations, and present data on predictive validity. Since the practice of substituting concurrent for predictive validity data is fairly common, the *Technical Recommendations* have specified that "Reports of concurrent validity should be so described that the reader will not regard them as establishing predictive validity."¹⁵

PREDICTIVE VALIDITY

Illustrative Uses of Tests for the Third Purpose—To Predict Future Performance

A person responsible for the selection of students likely to succeed in a given job, college, or curriculum is concerned with test scores as aids in doing a better job of selection. Counselors also use tests as predictors, but usually in placement rather than selection decisions. In some school situations, counselors recommend the placement of students in different ability groups; in such cases, they will often use data from more than one test. Counselors also are "predicting" (or helping the student to predict) from test data whenever scores on aptitude and interest tests are interpreted in terms of probable chances of succeeding in different colleges or in different vocations.

Primary grade teachers use reading readiness and intelligence tests as predictors when they use them as aids in grouping children into rapid-moving, slow-moving, and average groups for instruction in reading. Teachers use tests as predictors whenever they utilize test data in making decisions on the assignment of students to remedial or accelerated groups, or to the use of instructional materials that are either below or above grade level. All such judgments involve predictions regarding rate of progress or chances of success.

Test Scores as Predictors of Future Criterion Performance

The predictive validity of a test cannot be judged by an examination of its content. The basic procedure in studying the predictive validity of a

¹⁵ "Technical Recommendations for Psychological Tests and Diagnostic Techniques," *op. cit.*, pp. 201-238.

test is (1) to administer the test to a group of students or prospective employees, (2) follow them up and obtain data for each person on some criterion measure of his later success, and (3) compute a coefficient of correlation between individuals' test scores and their criterion scores, which may represent success in college, in a specific training program, or on the job. Such a coefficient of correlation may be called a predictive validity coefficient. We can interpret predictive validity coefficients in terms of the standard error of estimate¹⁶ of predicted scores. The formula for standard error of estimate for predicted criterion scores is as follows:

$$SE_{\text{criterion scores}} = SD_{\text{predictor}} \sqrt{1 - r^2}$$

where r is the predictive validity coefficient.

When we use this method of interpretation with typical validity coefficients (which usually range in size from .3 to .6), it seems that most predictor tests make little contribution to our accuracy of prediction. As an illustration, we will compute the standard error of estimate for a fairly high validity coefficient of .60 between a scholastic aptitude test and some criterion of success in a training program (such as grade-point-average).

If T -scores were used for both test and criterion, so that SD would be 10, we would obtain the following standard errors of estimate:

If $r = .60$

$$\begin{aligned} SE_{\text{criterion scores}} &= 10 \sqrt{1 - (.60)^2} = 10 \sqrt{1 - .36} \\ &= 10 \sqrt{.64} = 10 (.8) = 8.0 \end{aligned}$$

If $r = 0$

$$SE_{\text{criterion scores}} = 10 \sqrt{1 - 0} = 10 \sqrt{1.0} = 10.0$$

In other words, if we based our predictions on a test that had a validity coefficient of .60, rather than on one with no predictive validity, we would have reduced our standard error of estimate from 10 points to 8 points, or only 20 percent.

The formula for index of forecasting efficiency,¹⁷ which is based on this type of comparison, gives us a value of 20 percent. In other words, we reduce our error of prediction by only 20 percent when the validity coefficient is .60. Since we realize that predictive validity coefficients are seldom this high, the contribution of test scores to the prediction process seems very unpromising.

It is important, however, that we realize that this index is based on a comparison of predicted and actual scores, while our predictions in edu-

¹⁶ See Chapter 3, page 82.

¹⁷ The formula for the index of forecasting efficiency is as follows: $E = 100 (1 - \sqrt{1 - r^2})$. See J. P. Guilford, *Fundamental Statistics in Psychology and Education* (New York: McGraw-Hill Book Company, Inc., 1956), pp. 375-378.

cation and personnel work seldom require prediction of precise scores for individuals. We are usually satisfied with cruder predictions, such as the student's chance of achieving an acceptable rate in typewriting or short-hand, or his chances of making at least a C average in college.

The Use of Expectancy Tables in Interpreting Predictive Validity Data

Table 4.4 helps us to assess more realistically the value of a predictor test for helping us make the third type of judgment.

Table 4.4
Improvement in the Prediction of a Student's Chances for Success
When One Bases One's Predictions on a Test That Has a Correlation of
.50 or .60 with the Criterion Score and the Success Ratio is 50 percent^a

Student's standing on test		PREDICTED CHANCES OF SUCCESS VS. FAILURE WHEN		
		No information available to aid in prediction	Predictor tests used that correlate with criterion	
PERCENTILES	DECILE		$r = .50$	$r = .60$
90-99th	10	1 to 1	5 to 1	9 to 1
80-89th	9	1 to 1	3 to 1	4 to 1
70-79th	8	1 to 1	2 to 1	2 to 1
60-69th	7	1 to 1	1 to 1	2 to 1
50-59th	6	1 to 1	1 to 1	1 to 1
40-49th	5	1 to 1	1 to 1	1 to 1
30-39th	4	1 to 1	1 to 1	1 to 2
20-29th	3	1 to 1	1 to 2	1 to 2
10-19th	2	1 to 1	1 to 3	1 to 4
1-9th	1	1 to 1	1 to 5	1 to 9

Source: Adapted, with the permission of the publishers, from "Better than Chance," Test Service Bulletin No. 45 (New York: The Psychological Corporation, May, 1953), Table III. When a test user obtains a validity coefficient on the basis of local data of this type, he can construct tables similar to this one on the basis of (1) the validity coefficient and (2) the local percentage of successes and failures. Or he could use tables given in R. W. B. Jackson and A. J. Phillips, "Prediction Efficiencies by Deciles for Various Degrees of Relationship," Educational Research Series No. 11, Department of Educational Research (Ontario College of Education, University of Toronto, 1945).

^a It is assumed that one-half the students succeed on the criterion; for example, this table could be appropriately used for predicting "success in the freshman year of college" if one-half of students attain such success (as defined by a grade-point average of C or better).

Although Table 4.4 has more general application, let us assume that we are using an admission test of scholastic aptitude to predict a student's probability of making a C average in college when such an average is achieved by only one-half of the students. If we had no information about the student's scholastic aptitude, or if our test correlated .00 with the criterion, the best estimate for students in each decile (on scholastic aptitude) would be that one-half would "succeed" and one-half "fail" (with a C average or better being used as our definition of success). In other words, their chance of success vs. failure would be 1 to 1.

If we use a scholastic aptitude test with a validity coefficient of .50, the probability of success for students in each decile is given in the second column from the right. Here we see that a student's chance of success varies considerably with his score on the predictor test. The farther a student's score differs from the "indifferent zone" (the range of predictor scores where about one-half succeed and one-half fail on the criterion), the greater confidence we can have in our predictions. If a student's score on the predictor test is in the 10th or 9th deciles, his chances of success are 5 to 1, and 3 to 1, respectively. If he has scored in the lowest deciles (the 2d and 1st) on the scholastic aptitude test, his chances of success are much less than 50-50 (or 1 to 1); they are only 1 out of 3, and 1 out of 5, respectively. If we compare the two extremes on the scholastic aptitude test, we see that the chances of a person in the top decile succeeding is 25 times greater than that of a person in the lowest decile.

For students scoring in the top 20 percent, and the low 20 percent, the test scores are very helpful to us even though an r of .50 corresponds to an index of forecasting efficiency of only 13 percent. For the middle 40 percent of students this aptitude test has not increased our accuracy of prediction. For applicants with scholastic aptitude scores in this middle range, we are especially obligated to consider additional information when we make decisions about their admission. It might be desirable to administer another scholastic aptitude test to students in this group, a test especially selected (in terms of difficulty of items) to make differentiations *within* this middle group. On the other hand, if we are interested only in differentiating at the extremes, the present test would prove adequate, especially if we use supplementary information about grade-point average, reading comprehension, and the like to assist us in making judgments.

When we examine the last column of Table 4.4, we see that the seemingly modest increase in validity coefficient from .50 to .60 is reflected in considerably greater accuracy of prediction. The range of predictor scores, in which the test makes a negligible contribution to prediction, is reduced to the middle 20 percent of the group.

Correction of Validity Coefficients for Preselection

Validity coefficients of .50–.60 are typical for scholastic aptitude tests used at the high school level. Coefficients of this size, however, are seldom obtained when scholastic aptitude tests are correlated with grade-point average in college, especially if the college has high admission standards. Many of the applicants who would have obtained low criterion scores are eliminated in the selection process. That is, preselection of students results in a validity coefficient that is spuriously low.

Since predictions are to be made for the more heterogeneous group of *applicants*, the validity coefficient should be corrected for the homogeneity of the validation sample. For example, if the *SD* of the freshmen (on the scholastic aptitude test) is only .6 as large as that for the applicants, the validity coefficient should be corrected for “restriction of range”; for example, a validity coefficient of only .41 for the selected freshmen would be corrected to .60; this corrected coefficient is an estimate of the validity coefficient which would have been obtained if all applicants had been admitted.¹⁸ This coefficient would be more suitable than the uncorrected one to use in constructing a probability table like Table 4.4, for use with unselected applicants.

Importance of Obtaining Local Data Regarding Predictive Validity

When we are using tests to predict, it is very important that we obtain local validation data. In other words, we should have data that reflect local conditions, grading practices, and the like for our own group of employees or students. Test manuals should provide data about the predictive validity of the test for *typical uses and for different types of well-described validation groups*. Thus, on the basis of predictive validity data in test manuals, the test user can select one or more tests that seem promising for his local situation. The validity of the test for predicting grades or other local criterion data, however, is hypothetical until we verify the test’s predictive validity on the basis of local data.

Local expectancy tables are much more meaningful and acceptable than similar data from a test manual to those students, parents, or employees to whom test data are being interpreted in terms of probability of success. Local predictions should be based on the cumulation of considerable data, however, since the sampling error involved in computing correlation is sub-

¹⁸ A table for correcting values of r for “restriction in range” is given in most standard textbooks in statistics, for example, Quinn McNemar, *Psychological Statistics*, 3d ed. (New York: John Wiley and Sons, Inc., 1962). p. 144.

Table 4.5
Construction and Validation of Tests Used To Serve the Third Purpose—Predicting Future Performance

AIM IN TESTING (TYPE OF JUDGMENT TO BE MADE)

To predict a person's future performance on the test or on some external variable

EXAMPLES: To predict, from arithmetic test scores, students' grades in algebra; to predict, from eighth-grade scholastic aptitude test scores, how students will score in a similar examination given for college admission; to predict, from reading readiness scores, children's reading achievement (at the end of second grade).

TYPES OF TESTS USED FOR THIS PURPOSE

Scholastic aptitude tests, vocational aptitude tests (clerical, mechanical, and the like); special aptitude tests in fine arts; interest inventories; many special tests developed by personnel departments in business, industry, and the armed services to select and classify personnel; and the like.

GENERALIZED PROCEDURES FOR THE DEVELOPMENT OF TESTS FOR THIS PURPOSE

1. Study the nature of the *ultimate criterion* (that is, success on the job, in college, or in a specific course) to help in selecting or developing one or more *intermediate criteria* (which will later be used in validating test). If one is trying to predict successful performance on the job, a job analysis of the employees' work and a study of characteristics that differentiate successful and unsuccessful workers would be advisable.^a A careful study of criterion performance also helps in the developing of test items with predictive validity.
2. Devise a large number of test items on the basis of hypotheses about items likely to tap abilities related to success on the ultimate criterion (such as effective performance in a chosen career) or a significant intermediate criterion.
3. Administer the preliminary test to persons for whom criterion data will later be available—for example, applicants for employment, students applying for admission to college, students intending to enter the engineering profession.
4. As soon as the desired criterion data^b can be made available, obtain such data for examinees.
 - a. In many cases, one can obtain objective data on an intermediate criterion that seems crucial to success on the ultimate criterion.^c
 - b. If possible, avoid preselection in the validation group.^d
 - c. If supervisors' ratings, teachers' grades, or other subjective judgments are used as criterion scores, it is important that one avoid criterion contamination. Such contamination inevitably affects criterion ratings if the raters know how different applicants scored on the admission or employment tests.

Table 4.5 (Continued)
Construction and Validation of Tests Used To Serve the Third Purpose—Predicting Future Performance

5. Study examinee performance on *each item* of the predictor test in relation to criterion data.^a
6. Select items for the revised test that maximize the relationship between predictor test scores and criterion scores.^f
 - a. If a predictor test needs to be short (that is, if there is limited time for test administration), items should be selected that correlate high with criterion scores and low with other items. In this way we get maximum predictive value for the shorter time limit.
 - b. If a longer test is feasible, items that have high predictive value can be grouped into homogeneous, meaningful subtests. Then a specific school or a specific employer can, on the basis of local validation studies, assign the best weights to different subscores for specific purposes.^g

INFORMATION THAT SHOULD BE PROVIDED IN A TEST MANUAL REGARDING THE PREDICTIVE VALIDITY OF A TEST

1. Since a standardized predictor test will be used to predict criterion data in a variety of situations, the test author should make a number of validation studies—using more than one criterion for each of several validation groups.^h The validity coefficients should be obtained on a different group of subjects than the one used in item selection.
2. Sufficient data regarding validation groups should be given so that the user can judge whether his group resembles one of the validation groups sufficiently well that he can generalize their validity data to his own situation. If it appears that the test will be a valuable predictor in his local situation, he should test out this hypothesis as soon as possible by studying the relationship between predictor scores and criterion scores in his local situation.ⁱ
3. The test manual should provide data that enable the user to judge the reliability of the predictions he makes from predictor test data to criterion scores.^j
4. Unless the predictive validity of a test is approximately the same throughout the full range of scores, the standard error of estimate should be given for different score levels.^k
5. If predictive validity coefficients are corrected for unreliability of the criterion, the uncorrected coefficient should also be given so as not to mislead the prospective user concerning the effectiveness of the test in predicting actual criterion scores. Validity coefficients should *not* be corrected for the unreliability of the predictor test.

^a The "critical incident" approach is frequently used to help in developing hypotheses concerning "differences that matter." That is, one identifies, through ratings, production records, or some other means

persons who are considered to be highly successful and highly unsuccessful. Then supervisors are asked to report incidents that led to an employee's classification as a superior or inferior foreman, airplane

Table 4.5 (Continued)
Construction and Validation of Tests Used To Serve the Third Purpose—Predicting Future Performance

<p>pilot, or other type of employee for which the selection test is being devised. A similar study could be made of students showing a high or low level of attainment of an educational objective. This technique was developed by Flanagan and described in John C. Flanagan, "The Critical Incidents Technique," <i>Psychological Bulletin</i>, vol. 51 (July 1954), pp. 327-358.</p>	
<p>^b Thorndike and Hagen list the qualities desirable in a criterion measure, in order of importance, as follows: (1) relevance (or the extent to which score on criterion measure is affected by the same factors that determine over-all success on total job as worker or student in some special field); (2) freedom from bias (bias with respect to working conditions such as quality of equipment or sales potential of a geographic area, or bias with respect to variations in generosity of raters or grading standards of teachers); (3) reliability, and (4) availability (convenience, cost, length of time one has to wait for criterion data, and the like). Robert L. Thorndike and Elizabeth Hagen, <i>Measurement and Evaluation in Psychology and Education</i> (New York: John Wiley and Sons, Inc., 1961), p. 166.</p>	
<p>^c For example, grade-point average during the first year of engineering school constitutes a good intermediate criterion when the ultimate criterion is success in engineering; for unless one can be admitted to, and succeed in, the engineering curriculum, one will not have a chance to work in the profession. Shorthand dictation speed might or might not be a good predictor of the ultimate criterion of job success as a court reporter. Shorthand speed scores, although crucial in obtaining employment as a court reporter, might not correlate highly with success ratings within this highly skilled group. Hence admission to the profession of court reporting might be a far better criterion for a shorthand proficiency test than subjective ratings of success on the job. Amount of money earned might be a good</p>	
<p>criterion for sales work, provided that each salesman were assigned to an equally good territory; it would be a very poor criterion for success in teaching.</p>	<p>^d When a new test is being validated for college admission, employee selection, or some other selection situation, it is best to admit all persons in the validation sample to the college or to the job. Only in this way can one check on how effectively the test identifies persons who are likely to succeed or fail. If one cannot avoid preselection because of company or school policies, one must correct the spuriously low validity coefficients, computed on a selected group, for "restriction in range."</p> <p>^e For example, compare the performance on each test item for examinees who later succeed or fail in engineering school, have high or low production records on the job, and the like.</p> <p>^f For example, include in a revised engineering aptitude test only those items that show the highest correlation with freshman grade-point average in engineering school. Include in a revised reading readiness test only those items that show the highest correlation with children's scores on the second-grade reading achievement test.</p> <p>^g For example, if one had speed and accuracy scores in a test of type-writing ability, one would give relatively greater weight to speed in selecting individuals to type rough-draft copy from dictaphone records and relatively greater weight to accuracy in selecting employees to type perfect copy for photographic reproduction.</p> <p>^h For example, the author of a reading readiness test should report correlations between his test scores and reading achievement scores on two or more standardized tests for each of several validation groups.</p>

Table 4.5 (Continued)
Construction and Validation of Tests Used To Serve the Third Purpose—Predicting Future Performance

¹ Ebel stresses the need for making local validation studies of predictor tests. "For any user's group the test may be more or less valid than it was for the test author's tryout group. Quite possibly the user may even have a somewhat different purpose for testing than the test author had in mind. His criterion may be different. Again this means that the test may be more or less valid than the author reported. Under these conditions, how can a test author possibly publish fully adequate data on validity? The best he can do is to report validity under certain clearly specified and carefully restricted conditions of use. For the majority of possible uses of a test, validation becomes inevitably a responsibility of the test user." Robert L. Ebel, "Must All Tests Be Valid?" *The American Psychologist*, vol. 16 (Nov. 1961), p. 645.

² For example, expectancy tables like Table 4.4 might be included that indicate the range of possible criterion scores for a certain range in predictor-test scores. The standard error of estimate should be given for any suggested prediction formula for estimating future grade-point averages, production records, or other criterion scores. If predictions are made in terms of membership in categories, such as delinquent and nondelinquent, data should be provided that indicate the risk of misclassification.

³ For example, a test in arithmetic computation might have fairly high predictive validity for success in algebra within the range of low and average scores; however, increasing increments of accuracy in computation might not be associated with higher achievement in algebra.

stantial unless predictor and criterion scores on at least 100 cases and preferably 300 or more cases, are available.

Summary of Procedures Involved in Constructing and Validating Tests for the Third Purpose

Table 4.5 not only gives further examples of the use of tests for prediction but specifies the procedures that should be followed in developing a test designed to predict the future performance of individuals. In the last section of the table are listed the types of information that should be given in the manual of a published test so that the user can judge the probable predictive validity of the test for his own purposes. In making local validity studies, the suggestions regarding criterion data outlined in paragraph 4 (of the section on procedures for test development) should be followed.

CONSTRUCT VALIDITY

In our discussion of content validity, we focused our attention on achievement tests—tests that were intended to be representative samples of a universe of content or skills. In our consideration of concurrent and predictive validity, we were concerned chiefly with tests to be used in practical problems of selection and prediction—in estimating criterion data from test data. Some tests and inventories, however, are not samples of a defined universe nor are they designed to predict specific criteria. They presume to measure the degree to which individuals possess some trait or construct.¹⁹ Since tests that presume to measure the same trait frequently show low intercorrelations, we obviously cannot assume that test names accurately describe the dimension measured.

In appraising a test designed to measure a trait²⁰ or construct, we are

¹⁹ According to Cronbach and Meehl, a construct has three essential characteristics: (1) it is a postulated attribute assumed to be reflected in test performance, (2) it has predictive properties (persons who possess this attribute will in situation *X* act in manner *Y* with a stated probability), and (3) the meaning of each construct is given by the laws in which it occurs, with the result that clarity of knowledge of the construct is a positive function of that set of laws, termed a "nomological net." Lee J. Cronbach and Paul E. Meehl, "Construct Validity in Psychological Tests," *Psychological Bulletin*, vol. 52 (June 1955), pp. 281–302.

²⁰ The term "trait" is used in its broadest sense to include abilities, attitudes, personality dimensions, and the like. Cureton defines trait as follows: "When the item scores of a set of test-item performances correlate substantially and more or less uniformly with one another, the sum of the item scores (the summary score or test score) has been termed a quasi-measurement of 'whatever,' in the reaction-systems of the individuals, is evoked in common by the test items as presented in the test

concerned with all types of evidence that make the interpretation of test scores more meaningful, that help us to understand what the scores signify. "Construct validity is an analysis of the meaning of test scores in terms of psychological concepts."²¹ In their consideration of construct validity, the *Technical Recommendations* specify that "the manual should report all available information which will assist the user in determining what psychological attributes account for variance in test scores."²² In this sense of the word, construct validity includes the three other types. In many test-use situations, we do not care whether an achievement test measures a single ability or some unknown combination of several abilities, provided that the "test as a whole" fairly represents the universe we are sampling. Similarly, we do not care whether a group intelligence test measures an unknown combination of several abilities if it correlates highly with the Stanford-Binet or if it predicts a significant criterion, such as grade-point average.

For certain purposes, however, such as the testing of hypotheses through research, we would rather measure purer, single-factor traits. However, since many interrelated factors affect human behavior, we cannot find correspondingly pure criteria. For example, we might attempt to study the construct of "social introversion" or shyness. Some persons who rank high on this "construct" may actually appear unsociable; others may engage in an average number of social activities but may enjoy them less and experience more emotional stress in doing so. Hence, scores in any test of "social introversion" would not be consistently associated with, or highly correlated with, any single criterion. However, we might be able to make several hypotheses about ways in which individuals who are socially retiring would differ from those who are not—with respect to types of occupations pursued, leadership or followership roles assumed, symptoms of emotional stress when engaged in social activities, or behavior in experimental situations that allow opportunities to measure suggestibility, initiative in a leaderless group discussion, and other factors presumably related to the construct.

Several criteria, rather than a single criterion, are used. No single measure has status as *the* criterion for the trait. When the differences between

situation. This 'whatever' may be termed a 'trait.' The existence of the 'trait' is demonstrated by the fact that the item scores possess some considerable degree of homogeneity; that is, they measure in some substantial degree the same thing. We term this 'thing' the trait." Edward E. Cureton, "Validity," in E. F. Lindquist, ed., *Educational Measurement* (Washington, D.C.: American Council on Education, 1951), p. 648.

²¹ Lee J. Cronbach and Paul E. Meehl, "Construct Validity in Psychological Tests," *Psychological Bulletin*, vol. 52 (June 1955), pp. 281-302.

²² "Technical Recommendations for Psychological Tests and Diagnostic Techniques," *op. cit.*, p. 227.

high-scoring and low-scoring groups are in the predicted direction, the findings support *both* the construct validity of the test and our theory concerning the nature of the trait and its correlates in actual behavior. When the differences are not in the predicted direction, further evidence is needed before we can interpret whether we should question (1) the validity of the test, (2) our theories about the nature of the trait and related behavior, or both.

An Illustration of a Standardized Test Designed to Serve the Fourth Purpose

Many scholastic aptitude tests are designed to serve the fourth purpose (the measurement of individual differences with respect to hypothesized traits). In appraising such a test of general intelligence, we are concerned with all four types of test validity.

We peruse the manual for information on the test's *content validity*—the author's definition of intelligence and the criteria he has used in including or excluding items. For example, Lorge and Thorndike state in the manual for their intelligence tests: "The tests are avowedly measures of abstract intelligence." In discussing the content validity of *Lorge-Thorndike Intelligence Tests*, the authors indicate that the following characteristics of their test items elicit behavior that they would describe as intelligent:

1. The tasks deal with abstract and general concepts.
2. In most cases, the tasks require the interpretation and use of symbols.
3. In large part, it is the relationships among concepts and symbols with which the examinee must deal.
4. The tasks require the examinee to be flexible in his basis for organizing concepts and symbols.
5. Experience must be used in new patterns.
6. *Power* in working with abstract materials is emphasized, rather than speed.²³

The authors clarify that there are many types of ability, classifiable as intelligent, which are not included; for example, the test makes no attempt to appraise the use of intelligence in social situations or in situations involving mechanical comprehension. If the test user is seeking a test that emphasizes abstract intelligence, this statement from the manual, plus examination of items, will help him to judge the wisdom of his choice.

Concurrent and *predictive validity* are also of concern to users of intelligence tests. The data on the concurrent and predictive validity of the Lorge-Thorndike tests provide evidence that high-scoring students behave differently than low-scoring students in situations requiring intelligent behavior; for example, high-scoring examinees earned better grades and

²³ Irving Lorge and Robert Thorndike, *Technical Manual, The Lorge-Thorndike Intelligence Tests* (Boston: Houghton Mifflin Company, 1962), p. 14.

made higher scores on the Stanford-Binet and on other tests of intelligence. Hence, we see that data classifiable under other types of validity support the construct validity of this test as a measure of "abstract intelligence" in that high test scores are associated with superior performance on several criteria of intelligent behavior.

Despite all the data presented on the first three types of validity, the test user needs additional information that will help him to judge whether the test is measuring the construct of "abstract intelligence." For example, he would like to know the effects of practice or special coaching on test scores. If practice has an immediate and sizable effect on test scores, one would doubt whether the test measures stable, underlying abilities. The Lorge-Thorndike manual reports no studies on coaching; studies of the effects of practice, however, revealed that retesting after a week results in no average gain for the verbal battery, but an average gain of eight IQ points in the nonverbal battery.

Further evidence concerning the construct validity of this test comes from a factor-analysis study of the subtests. The technical manuals of many tests include tables of factor loadings, which the test user needs to interpret if he is to judge the value of the test for his purposes. The problems of construct validity cannot be adequately considered without some understanding of this important method of making test scores more meaningful (through systematic study of the pattern of a test's correlations with other test data and/or nontest data, for the same individuals).

The Use of Factor Analysis in Studies of Construct Validity

Explanation of the mathematical procedures involved in factor analysis is clearly beyond the scope of this textbook.²⁴ We will attempt, however, to help the reader understand how factor loadings are obtained and interpreted.

In our discussion (in Chapter 3) of the comparatively low reliability of differences between test scores, we called attention to the fact that many tests overlap considerably in the abilities they measure. Test interpretation and test development are aided by a study of the extent to which tests

²⁴ For an introductory discussion of factor theory, and the use of factor analysis in studying the structure of human abilities, see Lee Cronbach, *Essentials of Psychological Testing* (New York: Harper & Row, Publishers, Inc., 1960, Chapter 9; Jum C. Nunnally, *Tests and Measurements: Assessment and Prediction* (New York: McGraw-Hill Book Company, Inc., 1959), Chapter 9. For a brief discussion of both factor theory and methods, see J. P. Guilford, *Psychometric Methods*, 2d ed. (New York: McGraw-Hill Book Company, Inc., 1954), Chapter 16, or Philip E. Vernon, *The Structure of Human Abilities* (New York: John Wiley and Sons, Inc., 1950), pp. 1-24.

measure overlapping traits and by an interpretation of what these overlapping traits or factors appear to consist.

As an oversimplified illustration, we might examine data concerning our local test of arithmetic and a standardized intelligence test, which we intend to use with it, in the counseling of eighth-grade students wishing to take algebra.

Reliability coefficient for arithmetic test	.84
Reliability coefficient for standardized intelligence test	.94
Correlation between the two tests	.60

The reliability coefficient for the arithmetic test indicates that 84 percent of the variance in scores is "true variance," while 16 percent of the variance is attributable to *error variance*. If we square the r of .60 between the two tests, we find that 36 percent of the variance of scores on either test can be interpreted as representing overlapping abilities, *common* to the two tests.²⁵ This leaves 48 percent of the variance of scores on the arithmetic test that is attributable to *specific abilities*, that is, abilities not measured by the intelligence test.

Similarly, the variance of intelligence test scores may be interpreted as attributable to 6 percent, error variance; 36 percent, variance due to common or overlapping abilities; and 58 percent, abilities specific to the intelligence test, that is, not measured by the other test.

Factor analysis is a systematic procedure for studying the interrelationships between tests or other measures. Although a factor analysis cannot be performed with two tests, and should usually not be attempted unless we have ten or more carefully selected tests, this example gives some notion of what is meant by overlapping or common abilities. Certainly, in this case, we would need correlations with other tests before we could possibly interpret or name what "factors" are common to these two tests, which were not designed to measure the same abilities.

A factor analysis study always begins with a complete table of intercorrelations among a set of tests, in which each r appears twice (Table 4.6). Such a table is called a *correlation matrix*. A crude approach to factor analysis can sometimes be made by inspecting such a correlation matrix for groups or clusters of variables that show fairly high r 's with each other. In Table 4.6, we can locate the highest correlation coefficient,

²⁵ See the explanation in Chapter 3, page 82, that r^2 reflects the percent of variance in one variable that is explainable or predictable in terms of variance in the other; for example, in a situation of the type we are discussing, a predictive validity coefficient of .60 would indicate that 36 percent of the variance in criterion scores is predictable from variance in predictor-test scores. If all persons had the same predictor scores, the variance with respect to criterion scores would be reduced by 36 percent.

which is .74 between tests A and E. We then examine the other r 's with tests A and E to see if any other test has a substantial r with both of them. Test C obviously qualifies for a place in this cluster since its r 's with A and E are .63 and .57 respectively. Examination of the r 's for tests B, D, and F shows that they do not belong to this first cluster but that they show high intercorrelations with each other; the three r 's are .41, .48, and .58.

Table 4.6
A Hypothetical Correlation Matrix Showing All Intercorrelations
among Six Tests^a

ORIGINAL CORRELATION MATRIX						
	A	B	C	D	E	F
A	()	.02	.63	.05	.74	.10
B	.02	()	.03	.41	.09	.58
C	.63	.03	()	.02	.57	.09
D	.05	.41	.02	()	.12	.48
E	.74	.09	.57	.12	()	.05
F	.10	.58	.09	.48	.05	()

THE SAME CORRELATION MATRIX WITH COLUMNS AND ROWS REARRANGED TO IDENTIFY TWO CLUSTERS OF INTERRELATED VARIABLES						
	A	E	C	D	B	F
A	()	.74	.63	.05	.02	.10
E	.74	()	.57	.12	.09	.05
C	.63	.57	()	.02	.03	.09

	D	B	F
D	()	.41	.48
B	.41	()	.58
F	.48	.58	()

^a To find the correlation coefficient between any two tests (for example, A and F) look down the A column to find an r of .10 in the F row; or look in the A row and find the same value of r listed in the F column. Each correlation appears twice in the correlation matrix. The diagonal cells represent the correlation of each test with itself; with perfect reliability, each of these r 's would be 1.00; however, different values are used at different steps in the process of factor analysis.

The two clusters are more easily seen in the bottom half of Table 4.6 where we have rearranged the r 's so that the variables in the first cluster are in the first three columns and rows. Because of a clear-cut pattern of

interrelationships, which is rarely found in actual research studies, we have identified two clusters of tests that seem to represent common abilities or factors. If we could then study tests A, E, and F to see what psychological processes they seem to have in common, as well as the ways in which they differ from the other tests, with which they show low r 's, we might be able to suggest a name for an ability or trait that the tests in the cluster seem to be measuring in common, which is absent from (or greatly diminished in) tests outside the cluster.

However, clusters of tests are rarely as clear as in the hypothetical set of r 's in Table 4.6. Inspecting a large table of correlations for possible clusters is difficult and uncertain. The mathematical procedures of factor analysis surpass those of cluster analysis in several ways. Factor analysis involves less subjectivity, achieves more meaningful solutions, and gives greater assistance in clarifying what each test measures.²⁶

Every factor analysis study ends with a factor matrix, such as the one shown in Table 4.7 for the *SRA Achievement Series*, which shows the factor loading²⁷ or weight of each of the factors in each of the tests. Since a factor loading represents the correlation of the factor with the test, we can square the factor loading to obtain the proportion of variance in the test attributable to each factor. For example, in Table 4.7, each factor loading for the grammatical usage test can be squared and the variance in scores on this test interpreted as follows: squaring the factor loading of .72 gives .52, or 52 percent of the variance attributable to factor I; similarly, squaring the factor loading of .41 gives 17 percent of the variance attributable to factor III; negligible amounts are attributable to the other factors. The factor loading of a test on the chief factor it was designed to measure is called its *factorial validity*.

²⁶ Mathematical procedures for cluster analysis have been developed; but ordinarily their use is confined to *preliminary* study of a large correlation matrix. Cluster analysis provides a less economical explanation of the data in that there are generally many more clusters than factors. Moreover, cluster analysis classifies each variable or test as a unit, whereas factor analysis can assign different portions of the variance of each test to different factors. For a comparison of cluster and factor analysis, see Benjamin Fruchter, *An Introduction to Factor Analysis* (Princeton, N. J.: D. Van Nostrand Company, Inc., 1954), Chapter 2. Raymond B. Cattell, *Factor Analysis* (New York: Harper & Row, Publishers, Inc., 1952), Chapter 2.

²⁷ Through the mathematical procedures of factor analysis, the researcher can estimate the correlation of each test with each factor. These estimated correlations are known as *factor loadings*. The square of the factor loading for factor I on a specific test, for example, grammatical usage, indicates the proportion of the variance in test scores that is attributable to examinee differences with respect to factor I. If the examinees had identical scores on factor I, the variance in scores on this specific test would be reduced by that proportion.

Table 4.7
SRA Achievement Series, Battery for Grades 2-4
Factor Loadings for Grade 2 (300 Cases)

	Rotated Orthogonal Factor Loadings ^a				
	I	II	III	IV	V
Comprehension	.85	.15	-.03	.05	.14
Vocabulary	.85	.28	.02	-.03	-.04
Capitalization and Punctuation	.60	-.02	.40	.05	.00
Grammatical Usage	.72	.02	.41	-.10	.07
Arithmetic Reasoning	.78	.25	.02	.21	.00
Arithmetic Concepts	.68	.29	-.03	.36	.10
Computation	.46	.13	.06	.40	.00
Auditory Discrimination	.50	.20	.00	.00	.50
Visual Discrimination	.38	.20	.00	.30	.33
Sight Vocabulary	.78	-.11	-.08	.10	.48

Source: Adapted with the permission of the publisher from Louis P. Thorpe, D. Welty Lefever, and Robert A. Naslund, *SRA Achievement Series, Technical Supplement* (Chicago: Science Research Associates, Inc., 1957), Table 19, p. 25.

^a Factor I—the general achievement factor. "This factor occurs because the various tests measure, in part, the same cognitive skills. These skills, which are common to all cognitive-type tasks, are assumed to be almost the same as skills that are measured by the general factor in intelligence tests."

Factor II—the symbolic language factor. Each of the tests with significant loadings in this factor deals with such symbolic processes as abstracting, interpreting, relating, deducing.

Factor III—is composed of the two language arts tests; hence can be called the structural language factor, since it measures the knowledge of rules about the structure of language.

Factor IV—is labeled quantitative accuracy and principles. It is made up of the three arithmetic tests.

Factor V—at the second-grade level consists essentially of the three language perception tests.

To facilitate the work of factor analysts in the area of aptitude testing, the Educational Testing Service has developed a kit of reference tests for cognitive factors. These tests have been selected as having high factorial validity. The original 1954 kit has recently been revised; the 1963 kit includes 74 tests of 24 aptitude and achievement factors. Now that computers can do most of the drudgery of factor analysis, a number of very comprehensive, well-designed studies are being made. While most of the early studies involved 10 to 20 variables, it is not uncommon for recent

studies to involve 60 variables, with less work and time being required than in the pioneer studies.²⁸

As factor analysis studies become more extensive and more carefully designed, their findings will increasingly converge, confirming the existence of many factors. Fruchter includes a list of factors that have been verified in three or more studies.²⁹ We are gradually establishing a reference system of factors, in terms of which different tests can be described. We must, however, guard against an interpretation of factors as underlying dimensions of ability or temperament, which inevitably unfold from genetic potential and will be found in the same patterns in all cultures. Factors are useful, meaningful dimensions; they arise from the interaction of genetic factors with patterns of environmental factors that are interrelated in a specific cultural setting.

The "factors" of factor analysis really correspond to sets of responses whose joint occurrences are conspicuous in comparison to those of other possible sets of responses. . . . When we try to "interpret" a factor, therefore, we should attempt to consider the source of the frequencies of co-occurrence.

There are many possibilities. Among these are: (1) the learning of the performance of one response is prerequisite to, or implied in, the performance of another . . . (2) the learned behavior represented in one response has transferred to the other response; and (3) because of the accidents of personal history or because of the common experience of certain numbers of the group of persons, there was a higher probability that both of any pairs of responses were learned together than that either would have been learned alone.³⁰

All three of the possibilities listed above would help to account for the fact that individuals may score high or low in the numerical factor, without necessarily having a comparable score in the verbal factor of mental ability.

Fortunately, computers have also made it possible to perform factor analysis studies of *items*, with each item being considered as a variable. An author can develop or select a large number of items that seem to measure certain personality traits and determine through factor analysis

²⁸ Raymond B. Cattell, *Factor Analysis* (New York: Harper & Row, Publishers, Inc., 1952), p. 386.

²⁹ Benjamin Fruchter, *Introduction to Factor Analysis* (Princeton, N. J.: D. Van Nostrand Company, Inc., 1954), pp. 197-198, citing J. W. French, ed., *Conference on Factorial Studies of Aptitude and Personality Measures* (Princeton, N. J.: Educational Testing Service, 1952), p. 12.

³⁰ John B. Carroll, "Factors of Verbal Achievement," *Invitational Conference on Testing Problems* (Princeton, N. J.: Educational Testing Service, 1961), pp. 12-13.

the items that should logically be grouped together into subtests. Such factor analyses of items have been made on currently used personality inventories, with disconcerting results.³¹

The Concepts of Convergent and Discriminant Validity as Aspects of Trait Validity

Campbell³² emphasizes that the validity of a proposed trait or construct (such as "social introversion") should be carefully studied to see if the trait is distinguishable from other traits and whether two or more measures of the trait (by independent methods) tend to agree.

Studies should also be made to determine whether individual differences in trait scores are largely attributable to response tendencies³³ (such as tendency to agree with generalizations, to admit symptoms, or to answer questions in such a way as to create a good impression). For example, a great deal of research work was done with the *California F Scale* (designed as a measure of authoritarianism) before its trait validity was adequately studied. It was unfortunate that many studies were made concerning the relationship of "authoritarianism" to "rigidity" and other traits before it was found that a response tendency to accept overgeneralizations was one of the major factors in the "authoritarianism" scores; in fact, persons who agreed with the extremely worded, cliché-ridden statements of the *California F Scale* tended also to agree with their reversals.³⁴

Each test, rating scale, or other measure is a trait-method unit; and some of the variance in scores is due to method factors that are unrelated

³¹ For example, studies of the *Minnesota Multiphasic Inventory* by Comrey and Soufi, and of the *Edwards Personal Preference Schedule* by Levonian *et al.*, have indicated that the present grouping of items into subscales on these two tests departs markedly from the grouping that appears to be desirable on the basis of factor analysis studies. Andrew L. Comrey and A. Soufi, "Further Investigation of Some Factors Found in MMPI Items," *Educational and Psychological Measurement*, vol. 20 (Winter 1960), pp. 777-786; E. Levonian *et al.*, "A Statistical Evaluation of the Edwards Personal Preference Schedule," *Journal of Applied Psychology*, vol. 43 (December 1959), pp. 355-359.

³² Donald T. Campbell, "Recommendations for APA Test Standards Regarding Construct, Trait, or Discriminant Validity," *The American Psychologist*, vol. 15 (August 1960), pp. 546-553.

³³ A response tendency may be defined as "any tendency causing a person consistently to give different responses to test items than he would when the same content is presented in different form." Lee J. Cronbach, "Response Sets and Test Validity," *Educational and Psychological Measurement*, vol. 10 (Spring 1950), pp. 3-31.

³⁴ H. E. Titus and E. P. Hollander, "The California F Scale in Psychological Research: 1950-1955," *Psychological Bulletin*, vol. 54 (January 1957), pp. 47-64.

to the trait itself. For example, Thorndike³⁵ found that ratings of teachers' voices correlated .63 with ratings of their intelligence. He realized that this correlation must greatly overestimate the relationship between intelligence and pleasing quality of voice; that the correlation was largely due to the "halo effect" of the *rating method*. In other words, the rater who perceived a certain teacher as a superior teacher tended to rate this teacher high in these and other traits.

In studies of both concurrent and predictive validity, fairly high correlation coefficients, representing agreement or convergence, were sought as evidence of validity. However, the construct validity of a test (as a measure of a hypothesized trait) requires evidence of both *convergent* and *discriminant* validity. A test of a trait should show *convergent validity* (that is, fairly high *r*'s with other tests of the same trait and with measures of behavior that should be associated with it). A test of a trait should also show *discriminant validity*—low correlations with tests from which it is supposed to differ. For example, several "moral knowledge" tests developed during the 1920s were found to correlate more highly with intelligence tests than they did with each other.³⁶ A test of moral knowledge has inadequate trait or construct validity if it correlates higher with tests of intelligence and reading ability than it does with other tests of moral knowledge.

Campbell recommends the use of the multitrait-multimethod matrix³⁷ as an ideal first approach to studying the construct validity of a new test or rating scale. An example of such a matrix is given in Table 4.8. Measures of the same trait should correlate higher with each other than with measures of different traits involving the same method. Through a table of this type, one can study the *convergent validity* between independent measures of the same trait and the *discriminant validity* between measures of different traits.

Let us assume that when the committee was working on problem 5, the selection of gifted students, they decided to obtain ratings for students on three different traits: (1) seriousness of purpose; (2) responsibility; and (3) originality. For each of these traits, the committee developed a list of behaviors that would help to define the trait for teachers and student raters.

Since one of the teacher members of the committee was willing to make

³⁵ E. L. Thorndike, "A Constant Error in Psychological Ratings," *Journal of Applied Psychology*, vol. 4 (March 1920), pp. 25–29.

³⁶ Campbell, *op. cit.*, p. 548.

³⁷ A multitrait-multimethod matrix is a matrix of intercorrelations among tests representing at least two traits, each measured by at least two methods. An example is given in Table 4.8. Note that each coefficient appears only once, rather than twice as in the correlation matrix used in factor analysis.

a fairly elaborate study as the basis for his master's thesis, three methods of obtaining information on each of the traits were used:

METHOD 1. TEACHER RATINGS Each student was rated independently on a 5-point rating scale on "seriousness of purpose" and the other traits by the three sixth-grade teachers who had worked with him on the teaching-team program. The average rating for each pupil on each trait was translated into a stanine score.

METHOD 2. TEACHER NOMINATIONS Each teacher was asked to nominate 20-25 percent of his students whom he would describe as outstanding with respect to "seriousness of purpose" and each of the other traits. He was also asked to nominate 20-25 percent whom he would describe as least adequate in this respect. To obtain the raw score for each pupil the number of negative mentions was subtracted from the number of positive mentions. Pupils not mentioned were given a raw score of 0. Again, raw scores were converted into stanine scores.

METHOD 3. PEER-GROUP NOMINATIONS During the last month of the sixth grade, each sixth-grade pupil was asked to nominate his classmates for each of the traits, following the nomination procedure described under method 2. Although the raw scores were much larger under method 3 (because of the many peer ratings involved), the use of stanine scores made the scores comparable in size to those obtained under method 2.

In Table 4.8 are presented the intercorrelations of scores obtained on three traits by each of three methods. In the column headed A_1 , for example, are the correlations between scores on trait A by method 1 (teacher rating) and scores on all the other trait-method units. A careful study of the explanation of symbols (at the foot of the table) is necessary to interpret the table in terms of *convergent* and *discriminant* validity.³⁸ (The term "test" is used throughout even though ratings are involved. The procedures involved in using this type of matrix to study construct validity are the same, whether tests or ratings are involved.)

³⁸ A check should first be made to see whether each r is significantly different from zero. Tables for checking on the statistical significance of r 's are given in standard textbooks on statistics, for example, J. P. Guilford, *Fundamental Statistics in Psychology and Education* 3d ed. (New York: McGraw-Hill Book Company, Inc., 1956), p. 539. If data for one hundred students or more are involved, all the r 's in Table 4.8 would be significant at the 1 percent level. That is, there is only one chance in 100 that an r as large or larger than .20 would be obtained if the "true r " were zero.

Table 4.8
An Illustrative Multitrait-Multimethod^a Matrix (Table of Intercorrelations)

	Method 1: teacher ratings on traits			Method 2: teacher- nomination scores on traits			Method 3: peer-group nomination scores on traits		
	A ₁	B ₁	C ₁	A ₂	B ₂	C ₂	A ₃	B ₃	C ₃
METHOD 1: TEACHER RATINGS ON									
A ₁ Seriousness of purpose	(.81)								
B ₁ Responsibility	.70	(.84)							
C ₁ Originality	.63	.60	(.75)						
METHOD 2: TEACHER NOMINATION SCORES ON									
A ₂ Seriousness of purpose	.60	.63	.51	(.75)					
B ₂ Responsibility	.64	.75	.40	.62	(.77)				
C ₂ Originality	.54	.43	.54	.64	.44	(.70)			
METHOD 3: PEER-GROUP NOMINATION SCORES ON									
A ₃ Seriousness of purpose	.48	.52	.29	.40	.45	.34	(.85)		
B ₃ Responsibility	.56	.64	.31	.46	.49	.28	.75	(.75)	
C ₃ Originality	.30	.28	.41	.28	.30	.41	.35	.41	(.58)

Triangles drawn with solid lines enclose *r*'s involving same method, different trait. These triangles and the adjacent reliability coefficients make up same-method blocks.

Triangles drawn with broken lines enclose *r*'s involving different method, different trait.

^a Reliability coefficients (same method-same trait) are shown in the center diagonal in parentheses. Validity coefficients (same trait, different methods) are shown in boldface figures in the two other diagonals. This technique of analyzing a set of intercorrelations for evidence of convergent and discriminant validity is presented in Donald T. Campbell and Donald W. Fiske, "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin*, vol. 56 (March 1959), pp. 81-105.

The guide lines for interpreting data from such a multitrait-multimethod matrix are as follows:

CONVERGENT VALIDITY 1. The validity coefficients (boldface figures) should be sufficiently large to justify further research on the construct. This condition seems to be satisfied for all three traits.

DISCRIMINANT VALIDITY 2. Each validity coefficient (boldface) should be higher than the values *in its column and row in the triangles drawn with broken lines*; that is, a validity coefficient should be higher than r 's obtained between that trait and any other variable having neither trait nor method in common. A check of Table 4.8 reveals that the validity coefficients for trait A do not meet this standard, but that those for traits B and C almost meet it.

3. Each "test" (trait-method unit) should correlate higher with other tests measuring the same trait than with tests involving the same method and designed to measure different traits. Examination of Table 4.8 reveals that this standard is not even approximated. The r 's in the same-method triangles have a higher average than do the validity coefficients. Especially high, for example, is the r of .70 between teacher ratings on "seriousness of purpose" and "responsibility" and also the r of .75 between peer ratings on these two traits. Method variance equals or exceeds trait variance in the instruments studied. In other words, such method factors as the "halo effect" in ratings account for much of the convergent validity reflected in the validity coefficients.

Summary of Procedures Involved in Constructing and Validating Tests for the Fourth Purpose

A study of Table 4.9 helps the reader to realize that many of the tests used in guidance and research are measures of traits or constructs. Hence, no single criterion can be used in test development and test validation. Many interrelated research studies are needed, and test scores become more meaningful as more and more of the necessary data are cumulated and interpreted.

SUMMARY STATEMENT

As the student has already discovered, validity is a fundamental and crucially important concern in any type of measurement. A test does not have validity in general, but only in terms of its use for specific purposes and with specific groups. Hence, the subject of validity is the most complex one in the entire field of measurement.

The term "validity" refers to the value of a test as a basis for making judgments about examinees. For a test to be valid, it must measure "something"

Table 4.9
Construction and Validation of Tests Used To Serve the Fourth Purpose—Making Inferences Regarding Individual Status on a Postulated Trait or Construct

AIM IN TESTING (OR JUDGMENT TO BE MADE)

To make inferences concerning the degree to which a person possesses a hypothetical trait or construct, presumably reflected in the test performance

EXAMPLES: To make inferences from test scores concerning a person's anxiety-proneness, his level of motivation to achieve, or his suggestibility; to describe a person in terms of his pattern of abilities, personality traits, or interests.

TYPES OF TESTS USED FOR THIS PURPOSE

Tests of general intelligence or creative thinking abilities, personality inventories, or interest inventories (when used to describe persons rather than to predict specific criterion scores).

GENERALIZED PROCEDURES FOR THE DEVELOPMENT OF TESTS FOR THIS PURPOSE

1. Devise a large number of items, presumed to measure the trait, on the basis of
 - a. Hypotheses about the nature of the trait.
 - b. Empirical data on differential responses to items by groups that should differ with respect to the trait (as age groups on intelligence test items).
2. Through care in item writing and revision, minimize factors that would contaminate the assessment of the trait, for example
 - a. Make sure that the items do not require above-average reading ability or vocabulary, or knowledge based on specialized experience.^a
 - b. Make such revisions in directions, content, and form of the test as will reduce invalid variance due to
 - (1) Lack of clarity in communicating to the examinee.^b
 - (2) Response tendencies.^c
 - c. Unless speed in performance is a valid aspect of the trait, time limits should be established that will allow all examinees to complete items.
3. Administer the revised test to a large group of examinees,^d who constitute a heterogeneous group with respect to the trait being studied.

Table 4.9 (Continued)
Construction and Validation of Tests Used To Serve the Fourth Purpose—Making Inferences Regarding Individual Status on a Postulated Trait or Construct

4. By means of a factor analysis of items (considering each test item as a variable), select those items for each trait that will maximize the homogeneity of the trait score.^e If a factor analysis of items is not feasible, use other methods for improving the homogeneity of each test.^f
5. Administer the revised test to other groups of examinees (representative of the groups for which the test was designed). Also obtain for the same examinees data on
 - a. Tests for which one would predict fairly high correlations. Within this group there should be some tests that represent independent methods of measuring the postulated trait, and others that measure variables which should be, according to psychological theory, related to the postulated trait.
 - b. Tests for which one would predict fairly low correlations. This group should include tests that measure probable sources of invalid variance, such as individual differences with respect to response tendencies.
6. Interpret the findings with respect to the test's reliability and the test's convergent and discriminant validity.^h
 - a. If, according to findings from the study outlined above, two or more methods of measuring a trait do not show fair agreement, it may be that (1) neither the new test or the other measure(s) is adequate for measuring the trait, (2) *either* the new or older method(s) is inadequate, or (3) the trait is not a functional unity. Further research to improve the test itself is required before inferences can be made from test scores concerning the examinees' relative status on the postulated trait.
 - b. If, however, the new test of the postulated trait yields scores that (1) show fair concurrent validity with one or more independent measures of the same trait, (2) show sufficient stability of scores, and (3) show sufficient freedom from invalid variance, then another series of studies should be made.
7. Make a series of studies in which the investigator formulates and tests hypotheses concerning the behavior of persons who score relatively high, and relatively low, on the postulated trait.
 - a. Experimental situations should be specially designed to see if the behavior of high-scoring examinees differs from that of low-scoring examinees in ways that verify hypotheses based on psychological theory.
 - b. If a hypothesis is confirmed, such evidence would support *both* the validity of the test and the validity of the theory. If a hypothesis is not confirmed, we need further evidence to help us in judging whether we should question the validity of (1) the postulated trait, (2) the psychological theory on which we based our hypotheses concerning relationships between the trait and relevant behavior, or both.

Table 4.9 (Continued)
Construction and Validation of Tests Used To Serve the Fourth Purpose—Making Inferences Regarding Individual Status on a Postulated Trait or Construct

INFORMATION THAT SHOULD BE PROVIDED IN A TEST MANUAL REGARDING THE CONSTRUCT VALIDITY OF A TEST

1. Evidence concerning the homogeneity or internal consistency of each test should be provided.
2. The correlations of the test with any published, generally accepted measures of the same attribute should be reported.
3. Information concerning differences in test performance for groups differing in age, sex, and other characteristics that might affect the attribute should be included.
4. If the test is given with a time limit, evidence concerning the effect of speed on test scores should be presented.
5. The manual should report all available data concerning tests of hypotheses about relationships between scores on the trait and external variables that (according to psychological theory), should have low, moderate, or high relationships with the trait.
6. Correlations of trait scores with scores on one or more intelligence tests should be reported. It should be demonstrated that the test correlates higher with each of several trait-appropriate criteria than does the intelligence test.
7. For a personality inventory (or other measure of the voluntary self-description type), procedures used to minimize the effects of social desirability on trait scores should be reported. It should be demonstrated that the new test correlates higher with each of several trait-appropriate criteria than does the general "social desirability" factor.
8. For a personality inventory (or other inventory of the voluntary self-description type), procedures used in test construction and scoring to minimize the effect of acquiescence (tendency to agree) and/or other response sets should be reported. It should be demonstrated that the new test correlates higher with each of several trait-appropriate criteria than do response set scores.
9. A multitrait-multimethod matrix (similar to Table 4.8) should be presented so that the user can examine
 - a. Evidence of convergent validity (correlations between independent measures of the same trait).
 - b. Evidence of discriminant validity (relationships involving different traits measured by the same method, and different traits measured by different methods).¹

¹ For example, in developing an interest inventory, make sure that the activities listed do not involve technical terms that some students would not understand; also make sure that the activities listed are

ones that most students would be acquainted with, so that differential experience in the various interest areas does not unduly affect relative scores in those areas.

Table 4.9 (Continued)
Construction and Validation of Tests Used To Serve the Fourth Purpose—Making Inferences Regarding Individual Status on a Postulated Trait or Construct

<p>^b For example, directions concerning omission of questions, interpretation of response categories such as "frequently" or "occasionally," and the like.</p>	<p>J. R. Wherry and B. J. Winer, "A Method of Factoring Large Numbers of Items," <i>Psychometrika</i>, vol. 18 (June 1953), pp. 161-179.</p>
<p>^c Separate response set scores might be devised, such as a score on "tendency to agree" or "tendency to choose responses that will create a good impression." If feasible, special methods (such as forced-choice questions) can be used to reduce the effects of response tendencies on test scores.</p>	<p>^f F. B. Davis, "Item Analysis in Relation to Educational and Psychological Testing," <i>Psychological Bulletin</i>, vol. 49 (March 1952), pp. 97-121.</p>
<p>^d In order for the results to be sufficiently stable, at least 300 examinees should be tested. Jum C. Nunnally, Jr., <i>Tests and Measurements: Assessment and Prediction</i> (New York: McGraw-Hill Book Company, Inc., 1959), p. 144.</p>	<p>^g For research concerning "social desirability" (the tendency to respond so as to create a good impression), see A. L. Edwards, <i>The Social Desirability Variable in Personality Assessment and Research</i> (New York: Holt, Rinehart and Winston, Inc., 1957).</p>
<p>^e Special methods have been developed for approaching a problem involving such a large number of intercorrelations, for example:</p>	<p>^h The terms "convergent" and "discriminant" validity are explained on page 138. ⁱ For additional illustrations for requirements 5 through 9 the student is referred to Donald T. Campbell, "Recommendations for APA Test Standards Regarding Construct, Trait, or Discriminant Validity," <i>The American Psychologist</i>, vol. 15 (August 1960), pp. 546-553.</p>

consistently (or with reliability), and that "something" must be either a representative sample of the behavior we wish to judge, or it must have demonstrated relevance to that behavior.

If we wish to know how individuals perform at present with respect to certain skills and knowledges, we try to devise a test which samples those skills and knowledges. If the behavior to be measured can be exactly defined (such as the correct spelling of the seventh-grade list of spelling words), *content validity* can be attained by including in the test, a random sampling of the complete list of spelling words, multiplication facts, or some other defined universe. Whenever the skills or knowledges to be sampled cannot be accurately defined and randomly sampled, human judgment must enter into the selection of learnings to be tested. As a guide for test construction, a table of specifications of the test content should be developed and followed. Our concern, however, is still with the representativeness of the test sample in terms of the universe about which we wish to make inferences.

If we use multiple-choice questions on spelling and arithmetic, or some other indirect procedure, for measuring the criterion behavior about which we wish to make judgments, we are obligated to assess the *concurrent validity* of our test, that is, the relationship of individuals' test scores to their results on some measure of criterion behavior external to our test. For the examples given above, the external criteria would be students' scores on dictation tests of spelling and on arithmetic tests involving actual computation. Or the concurrent validity of a group intelligence test might be studied by determining the relationship between students' scores on that test and their scores on an individual intelligence test, for which it was designed to be an economical substitute.

There are many situations in school and industrial work in which one wishes to predict a person's future performance in a subject, curriculum, school, or job. The *predictive validity* of a test cannot be adequately judged without following up a group of examinees to see how well they achieve on such criteria as job performance or average grades in specific college curricula. Predictive validity coefficients between test scores and appropriate criterion scores must be obtained. Locally developed expectancy tables can be used to aid counselors and other professional workers in estimating the probability of the future attainment of certain criterion scores, on the basis of the students' obtained test scores.

When a test presumes to measure the degree to which individuals possess some trait or construct, evidence concerning its *construct validity* must be obtained. Tests that claim to measure the same trait frequently show low inter-correlations; hence we cannot assume that test names accurately describe what is being measured. In many situations, we may not care whether the test name is accurate or whether the test measures a unitary, independent variable, provided the results serve some practical purpose, such as increasing the percentage of trainees who succeed in a training program. However, in research studies or in situations in which we wish to *describe* individuals as the basis for inferences regarding several decisions, we prefer to measure relatively pure traits which have meaning in terms of psychological concepts and which enable us to make inferences concerning correlated behaviors.

In studying the construct validity of a test which presumes to measure a relatively pure trait, several criteria must be used. For example, many hypotheses can be made concerning the differences in test and non-test behavior between groups of high-scoring and low-scoring students on a specific intelligence test. If correlations were obtained between scores on that test and several

other variables, psychological theory would predict high correlations with some variables and low or negative correlations with other variables. Confirmation of hypotheses based on theory would provide evidence for the construct validity of the test.

Factor analysis is used to study the extent to which tests measure unitary, independent dimensions, and also to help in interpreting the major sources of variation in test scores. The factor loading of a test on the chief factor it was designed to measure is called its *factorial validity*.

In order to assist the student in gaining a thorough understanding of this complex subject, the authors have prepared four summary tables, one for each of the aims listed in the "Technical Recommendations." In each table, we have given a generalized outline for the development of tests of the specified type. These procedures are not intended as a short course in the construction of standardized tests, but rather as an aid to the student in understanding the four types of validity and in comprehending what procedures should be used by test authors and publishers. They should also deter the student from amateur test construction, except in the development of tests for the first purpose, that is, measuring student achievement in knowledges and skills in the subjects he teaches. For example, if a student who is undertaking a research study for his Master's thesis decides to develop a test or inventory to measure some trait, an examination of Table 4.9 will help him to realize how far his procedures differ from those recommended, and how cautious he should be in making inferences from his test data.

In order not to obscure the major points in each of these tables, much of the illustrative material has been placed in the footnotes. The student is urged to read carefully all footnotes giving illustrative material, for such study will greatly increase his understanding of the generalizations listed in the tables.

SELECTED REFERENCES

- CAMPBELL, DONALD T., "Recommendations for APA Test Standards Regarding Construct, Trait, and Discriminant Validity," *American Psychologist*, vol. 15 (August 1960), pp. 546-553.
- CRONBACH, LEE J., "Validity," in C. W. Harris, ed., *Encyclopedia of Educational Research*. New York: The Macmillan Company, 1960.
- , AND PAUL E. MEEHL, "Construct Validity in Psychological Tests," *Psychological Bulletin*, vol. 52 (June 1955), pp. 281-302.
- EBEL, ROBERT L., "Obtaining and Reporting Evidence on Content Validity," *Educational and Psychological Measurement*, vol. 16 (Autumn 1956), pp. 269-282.
- FLANAGAN, JOHN C., "The Critical Incident Technique," *Psychological Bulletin*, vol. 51 (July 1954), pp. 327-357.
- HUDDLESTON, EDITH M., "Test Development on the Basis of Content Validity," *Educational and Psychological Measurement*, vol. 16 (Autumn 1956), pp. 283-293.
- KACZKOWSKI, HENRY R., "Using Expectancy Tables to Validate Test Procedures in High School," *Educational and Psychological Measurement*, vol. 19 (Winter 1959), pp. 675-677.
- LENNON, ROGER T., "Assumptions Underlying the Use of Content Validity," *Educational and Psychological Measurement*, vol. 16 (Autumn 1956), pp. 294-304.

- MCCABE, GEORGE E., "How Substantial is a Substantial Validity Coefficient?" *Personnel and Guidance Journal*, vol. 34 (February 1956), pp. 340-344.
- SUPER, DONALD E., AND JOHN O. CRITES, *Appraising Vocational Fitness*, rev. ed. New York: Harper & Row, Publishers, Inc., 1962, Chapter 3.
- WESMAN, ALEXANDER G., "Expectancy Tables—A Way of Interpreting Test Validity," *Test Service Bulletin* No. 38. New York: The Psychological Corporation, 1949. Available on request.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. Cite specific examples of ways in which test results can be used for each of the four purposes outlined by the American Psychological Association (page 106).
2. Examine a standardized test in your major subject field and evaluate its content validity.
3. Indicate whether concurrent or predictive validity would be studied in each of the following situations:
 - a. A school wishes to use a short form of an accepted reading readiness test.
 - b. A school wishes to select students for an accelerated class in algebra.
 - c. A college wishes to select freshmen who are likely to complete four years of college.

What would the criterion be in each of the three validation studies?

4. When will a teacher need to consider concurrent validity in selecting achievement tests? How is the concurrent validity of a test determined?
5. Administer a short multiple-choice test in arithmetic to a class. Then have the students take a traditional test involving these same problems. Show, by a comparison of ranks, the extent of agreement between students' scores on these two types of tests. What are the advantages and limitations of each approach? What type of validity is being studied?
6. Administer a standardized multiple-choice test of spelling to a class. Then dictate the same spelling words to the group. Show by comparison of ranks the extent of agreement between student scores on these two measures of spelling achievement. What are the advantages and limitations of each approach? What type of validity is being studied?
7. Select two silent reading tests designed for the same grades. Describe and compare them in terms of the skills and abilities measured.
8. Obtain manuals for two or more aptitude tests. What evidence on the predictive validity of these tests is presented? What criteria were used?
9. What are the advantages and limitations of the following validity criteria for appraising the predictive validity of a college entrance test of mental abilities? (a) grade-point-average during the freshman year, (b) graduation from college, (c) ratings by supervisors on success in postcollege vocations.
10. The correlation between entrance examination scores and grade-point-average in a law school is only .20. The average IQ of these students is 130. Explain why you would expect the correlation to be low.
11. Why are tests which presume to measure personality traits especially difficult to validate?

Application of the Principles of Measurement in the Selection of Tests

In the preceding chapters we have studied three basic areas of concern in measurement. The processes of norming tests, estimating their reliability, and studying their validity have all been presented and illustrated. From a study of these chapters, the student has discovered (1) that there is no best type of converted score, but that each type has advantages and limitations for specific purposes; (2) that there is no arbitrary standard for the size of a reliability coefficient and no method of computing reliability that is ideal for all purposes; and (3) that there can be no single basis for ranking tests in order of their over-all validity but that the validity of a test must be determined for each purpose for which test results will be used.

In this chapter, we will focus our attention on the selection of tests *for specific purposes*; we will present a test evaluation form that is designed to focus the user's attention on his purposes in testing; and we will indicate the sources of information available to help us in the location of suitable tests and in the process of their evaluation. Before we discuss test selection, however, we will consider (in the first chapter section) the different types of tests available for use.

TYPES OF TESTS AVAILABLE

All tests are designed to obtain samples of examinee behavior. Instead of waiting to observe behavior as it naturally occurs in the course of daily life, the test is designed to elicit from examinees the behavior the test user wishes to evaluate.

Standardized Tests

One of the obvious distinctions made among tests is that between teacher-made tests and standardized tests. The essential characteristic of a

standardized test is that the test materials, the procedures for administration, and the procedures for scoring have been so carefully developed that the test can be given and scored in the same manner at different times and places.

Only when a test has been so designed is it feasible to proceed with the norming of a test on representative samples of the population(s) with which we would like to compare our examinees. The norming process (that is, administering the test to representative age, grade, or occupational groups) is often called the standardization of the test; it should be evident, however, that the standardizing of test content, procedures, and scoring is actually a prerequisite to norming.

The term "standardized test" has come to signify a measuring instrument with the following characteristics:

1. Specific directions for administering the test are stated in detail, usually including even the exact words to be used by the examiner in giving instructions and specifying exact time limits. By following the directions, teachers and counselors in many schools can administer the test in essentially the same way.
2. Specific directions are provided for scoring. Usually a scoring key is supplied that reduces scoring to merely comparing the answers with the key; little or nothing is left to the judgment of the scorer. Sometimes carefully selected samples are provided with which a student's product is to be compared.
3. Norms are supplied to aid in interpreting the scores.
4. Information needed for judging the value of the test is provided. Before the test becomes available for purchase, research is conducted to study its reliability and validity.
5. A manual is supplied that explains the purposes and uses of the test, describes briefly how it was constructed, provides specific directions for administering, scoring, and interpreting results, contains tables of norms, and summarizes available research data on the test.

Classification of Tests According to Degree of Indirectness of Measurement

Most tests appear to be, and many are, inadequate substitutes for the direct study of behavior in everyday, real-life, situations. It is important that those persons who prefer to depend on direct observation in natural situations understand the obstacles to direct measurement. Only then can they accept the desirability of using indirect approaches that have demonstrated validity. On the other hand, it is only as the testing enthusiast faces up to the very natural tendency for tests to become several degrees removed from criterion behavior that he understands the limitations of paper-and-pencil tests and the need to obtain as much evidence as possible regarding the meaning of individual differences in test scores.

There are many obstacles to direct measurement of behavior in natural situations. Direct measurement is based on a representative sampling of

criterion behaviors in the real-life situations in which we are really interested. Indirect measurement involves the measurement (usually by methods that are more reliable and efficient) of behavior that is presumably related to the ultimate criterion behavior (for example, to desired changes in student behavior or successful performance in a chosen career). Almost all of our tests, ratings, and other evaluation techniques involve measurement that is indirect in some degree.

The widespread use of indirect measures may be largely attributed to several obstacles to direct measurement.

Direct measurement may be impossible because of the *delayed appearance* of the desired criterion behavior. For example, one of our major aims in teaching nutrition is to have future homemakers plan nutritious meals when they have families of their own; one of our major aims in civics instruction is that our students (when they become adults) should study thoroughly all issues on which they will vote. Obviously such future criterion behavior cannot be directly sampled, but only indirectly approached through the study of what students can do, or will do, in current situations.¹

Current criterion behavior may be *inaccessible, or not readily observed*. For example, one of our ultimate objectives in driver-training courses is to have students drive safely when they are not under adult supervision; one of our objectives in teaching science is to interest our students to the extent that they voluntarily read articles in newspapers and magazines about advancements in science. It is not convenient or practical for the teacher to make firsthand observations in these areas. Hence, he may substitute a questionnaire filled out in class, which he recognizes as a very poor substitute for direct observation.

Another problem is the *infrequency* of current occasions for observing the desired criterion behavior. For example, the instructor in first aid or lifesaving will probably witness no real-life situations in which the skills his students have learned must be demonstrated. One of the chief advantages of a test is the efficiency with which samples of behavior can be obtained.

Still another obstacle to direct measurement of criterion behavior is the *lack of comparability* of real-life situations from person to person. A visiting coach who tries to assess the players on a team realizes that a single game gives him an inadequate basis because players have had different opportunities to show their skills. He may be able to identify those who rank at either extreme; the ranking of other players would not be justified. If a

¹ Even if we did a follow-up study in an attempt to sample such criterion behaviors for students who had been in the nutrition or civics courses, as compared with those who had not, it would be almost impossible to infer that differences found were due to our instruction because of the many other factors that would influence delayed criterion behavior.

coach controls the situation so as to make conditions more nearly comparable among players, he has a better basis for comparing the players' ability to bat balls or catch fly balls. When he does so, however, he is taking some of the steps that are necessary in developing a test.

Perhaps the chief reason for the popularity of indirect measurement is that direct measurement of many types of criterion behavior is *very costly in time and effort, or so inefficient as to be impracticable*. For example, if we want to make inferences about students' spelling ability in their regular written work, we would have to examine thousands of words of writing for each pupil. Moreover, students try to avoid using words that they do not know how to spell. Using a test list of spelling words, pretested to eliminate those that do not differentiate between good and poor spellers at this grade level, would constitute a far more efficient approach. And, if we can show that the scores on a *recognition* spelling test correlate well with dictation-spelling scores, this still more indirect approach may be justifiable.

Another obstacle is the *complexity* of most criterion behavior. The complexity of the criterion behavior implied in "success as a teacher" is a good illustration. We could decide to observe the actual process of teaching; here the problems of obtaining a representative sampling of teaching situations, defining criteria for judgment, and assigning scores in terms of what we observe, appear formidable. If we decide that we are interested in the product of teaching, rather than the actual performance, we again face formidable problems in assessing all the significant aspects of student growth, some of which are much more easily measured than others.²

The various tests we use vary in the extent to which they approach direct measurement of criterion behavior.

The most direct is the "work-sample" or "identical-elements" test. The examinee is given special occasion to do some of the tasks on which we want to appraise his competency. In this group we could classify most tests of arithmetic computation, reading comprehension, typewriting, shorthand dictation, and the like. Many of the performance tests discussed in Chapter 12 are of the work-sample or "identical-elements" type.

A much more common type of test is the "related-behavior" type. The behavior elicited by such tests may or may not be similar to criterion behavior; the test behavior, however, is related to criterion behavior in the statistical sense, that is, test scores are substantially correlated with criterion scores. Tests of simulated driving behavior, used in many driver training courses, illustrate "related-behavior" tests in which test and criterion behavior are similar. So also would tests in which students demon-

² The preceding discussion regarding indirectness of measurement is largely based on a list of obstacles to direct measurement presented in E. F. Lindquist, "Preliminary Considerations in Objective Test Construction," *Educational Measurement* (Washington, D.C.: American Council on Education, 1951), pp. 143-146.

strate athletic skills under controlled, rather than team-play conditions. In other tests of related behavior, test behavior is not very similar to criterion behavior. Many tasks included in reading readiness tests are not obviously related to reading achievement. Nor would one suspect that scores on the *Minnesota Clerical Test* would show a substantial relationship to production records for packers, wrappers, and inspector-packers.³

A third basic test type is the "verbalized-behavior" type of test, in which behavior situations are described to the examinee and he tells how he would behave in those situations. The examinee may be required to indicate in his own words how he would plan an experiment, budget his money, or plan a nutritious meal. A nursing student, for example, might select the procedures he would use if a patient presented a specific pattern of symptoms.

A fourth type measures only the knowledge of facts and principles needed by the student in order to show the criterion behavior. We can find out through a test how much the student knows about a car and about traffic rules; how much he knows about the nutritional elements in different foods or about methods of first aid. The knowledge outcomes of education are important. Knowledge is necessary, but not sufficient, to the achievement of the ultimate objectives of education.

This classification of tests according to degree of indirectness in measurement is one of the most basic. Although there are no sharp lines of demarcation, the four types represent various degrees along a continuum of relevancy to the criterion behaviors, on which we wish to compare individuals.

Other Classifications of Tests on the Basis of Procedures

There are many other possible classifications of tests. Many of these are concerned with such procedural differences as:

1. Group tests (which can be administered to groups or to individuals) vs individual tests (which must be administered individually).⁴
2. Pencil-and-paper tests vs performance tests (the latter term usually being applied to tests requiring the use of physical objects and the application of physical and motor skills).
3. Speed tests vs power tests. In a speed test, the tasks presented are of approximately the same difficulty; administration time is limited so that none, or

³ M. L. Blum and B. Candee, "The Selection of Department Store Packers and Wrappers with the Aid of Certain Psychological Tests," *Journal of Applied Psychology*, vol. 25 (June 1941), pp. 291-299; E. E. Ghiselli, "Tests for the Selection of Inspector-Packers," *Journal of Applied Psychology*, vol. 26 (August 1942), pp. 468-476.

⁴ A comparison of group and individual tests will be made in Chapter 6 on aptitude testing.

almost none, of the examinees can finish; the score reflects the examinee's speed of reading, typing, proofreading or performing some other function. In a power test all, or almost all, examinees are given sufficient time to complete the test; the tasks are arranged in order of difficulty; the examinee's score reflects his accuracy and the level of difficulty at which he can successfully perform.

These differences in testing procedure are illustrated in Chapter 6 on aptitude testing.

Classification of Tests on the Basis of Content

The only other classification of tests that will be considered in this chapter is based on test content. The major distinction here is between *ability* tests and tests or inventories of *personality, interests, and attitudes*.

TESTS OF ABILITIES In tests of abilities

1. The goal is to measure the individual's *maximum performance*.
2. The examinee perceives the situation as one in which he should strive for accuracy and provide evidence of competency.
3. Comparison of results for different individuals is based on the assumption that all examinees are equally well motivated.
4. There are external standards of correctness on which experts agree. These external standards provide the basis for a scoring key, by which all answers can be evaluated.
5. The test author(s) attempt to reduce ambiguity of test content so that all persons will be working on essentially the same tasks.

Tests of ability can be grouped into two major categories:

1. aptitude tests, which are used to predict a person's future performance in some educational program or in some vocation (to be discussed in Chapter 6).
2. achievement tests that measure a person's present knowledge or level of performance in order to appraise individual or group success in past learning activities (to be discussed in Chapters 11–13).

Achievement tests are more heavily weighted with tasks that measure the students' learnings in specific courses, while aptitude tests include novel tasks and/or tasks with which all students are likely to have had previous experience (through learnings outside school and through a common core of required subjects).

Some tests are not easily classifiable as aptitude or achievement tests. For example, achievement tests that measure student's progress toward the over-all goals of the educational program, and scholastic aptitude tests that are designed to measure cognitive abilities developed through school experience, occupy an intermediate position. In Figure 5.1 we have at-

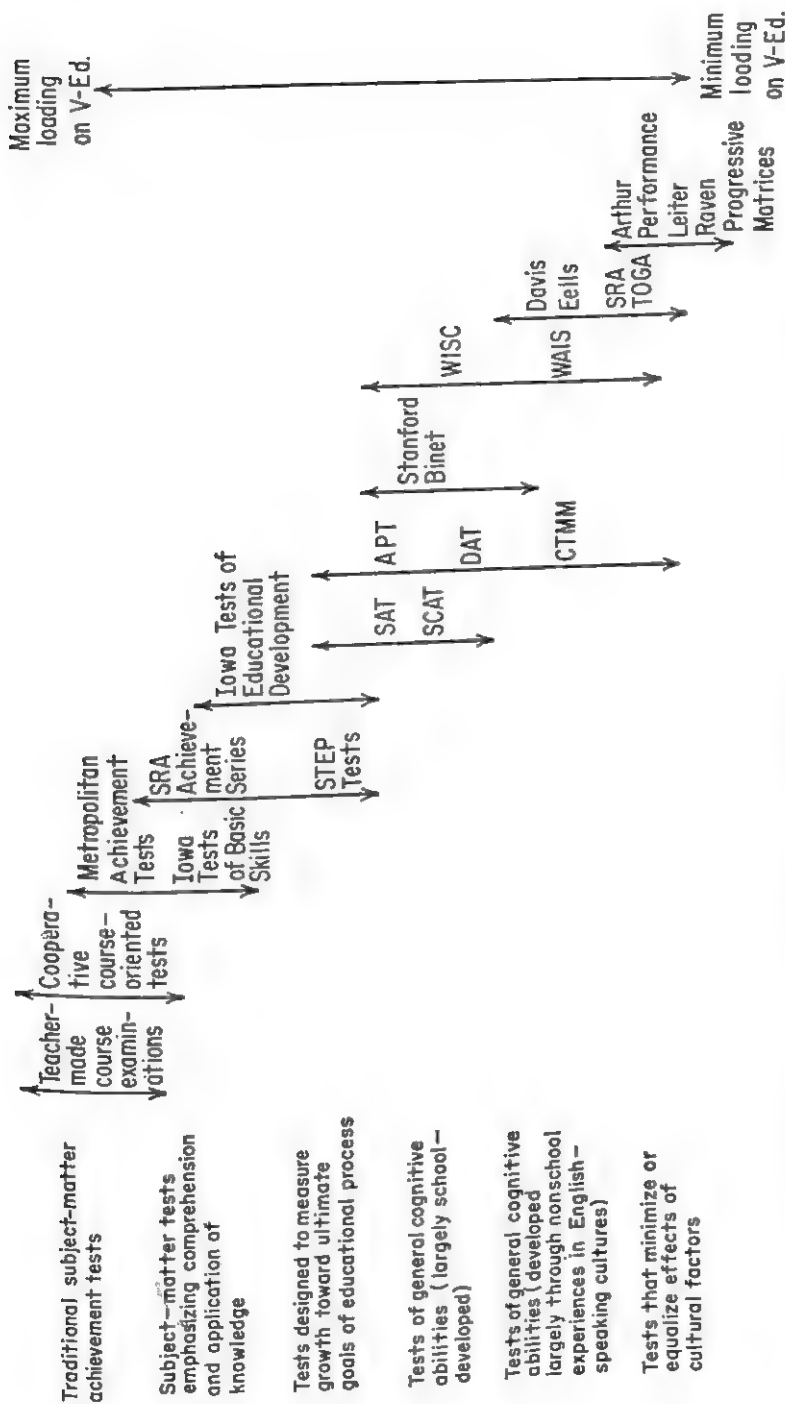


Fig. 5.1 Selected Ability Tests Representing Relatively Greater or Less Emphasis on the Verbal-Educational Factor

The development of this chart was suggested by similar charts on ability tests from Anne Anastasi, *Psychological Testing*, 2d ed. (New York: The Macmillan Company, 1961), p. 486; and Lee Cronbach, *Essentials of Psychological Testing* (New York: Harper & Row, Publishers, Inc., 1960), p. 235. For complete titles of tests and information concerning them, the reader is referred to the classified list of tests in the Appendix.

tempted to portray, through the use of a number of examples, the way in which different standardized tests represent varying amounts of emphasis on the verbal-educational factor.⁵

INVENTORIES OF PERSONALITY, INTERESTS, OR ATTITUDES Many tests in this area are really questionnaires involving self-report or self-description. In fact, the term "inventory" rather than "test" is being increasingly used.

In inventories of personality, interests, or attitudes (or in the observation and rating of these traits in real-life situations),

1. The goal is to sample the individual's *typical performance*,⁶ and he is encouraged to react in any way that is natural or typical for him.
2. Examinees can usually change their responses at will. Whereas an examinee cannot fake answers to an ability test, he can usually fake his replies to inventories of attitudes, interests, or personality traits. Every effort is made to have the individual perceive the testing situation as one in which he can feel free to show, or report, his typical behavior, feelings, and attitudes. In a counseling situation, the inventory is presented as an aid to counselor understanding and to self-understanding. In an employment situation, the technique is presented as an aid in maximizing the examinee's job satisfaction through placing him in a job best suited to his temperament or interests. Sometimes the purpose of the inventory is disguised, as when interest inventories are scored by personality-trait keys.
3. Interpretation of results for individuals is based on the assumption that all respondents are equally willing to report typical behavior; or provision is made to correct individual scores for tendencies toward deception and test-taking defensiveness.
4. There are usually no external standards of correctness, the responses being summarized in terms of categories.⁷

⁵ The term used by British factor analysts to represent competency in school-developed learnings.

⁶ It is recognized that test users vary in the degree to which they interpret responses to personality inventories as (1) having face value as honest and insightful responses, or (2) symptoms that have been shown empirically to be related to job success, job satisfaction, probability of improvement under therapy, or some other criterion. Even where responses to inventories are scored merely as symptoms, however, the assumption is made that examinees are responding to the questionnaire in a manner that is *typical* of them. In other words, the person who feels impelled to give many of the "good-impression" type of responses in an employment situation is assumed to show similar behavior in other situations in which he would feel vulnerable or on the defensive.

⁷ When an inventory is to be scored in terms of an individual's place on a desirable-undesirable continuum, such as neuroticism or predicted success in a specific job, external standards of "correctness" do exist. In such situations it is recognized that many examinees will strive for a maximum score; hence test items that are not easily faked are used and/or corrections are made for each individual's tendency to fake.

EVALUATING TESTS FOR USE FOR SPECIFIC PURPOSES

The principles of measurement (discussed in Chapters 2–4) become more meaningful when students have experience in applying them to the evaluation of specific tests for proposed uses. Ordinarily such experiences are most valuable if the student evaluates tests within subject areas in which he is teaching or plans to teach.^a

It is when the student evaluates tests in his own subject field that he can most effectively judge the content validity of tests. Moreover, ambiguities in the wording of items, poor selection of distractors (false alternatives), and other weaknesses in test items are most easily detected in one's own subject field. The prospective counselor should evaluate tests that will aid him in helping students to make wise decisions in their life planning.

Table 5.1
Summary of Data Needed in the Appraisal of a Standardized Test^a

REFERENCE DATA

1. Title _____
2. Names of major subtests: _____
3. Author(s) _____
4. Publisher _____
5. Range in grades _____
6. No. of forms _____
7. Purposes for which test is recommended by author: _____
8. Intended use (purpose and group for which test is being evaluated): _____

CONTENT VALIDITY (especially important for achievement tests)

9. Abilities and skills that this test is designed to sample.
10. Bases for selecting items (Sources of items and criteria for inclusion).
11. Your comments regarding the appropriateness of the item content for your local curriculum or for the specific purposes for which you will use test.

CONCURRENT AND PREDICTIVE VALIDITY (especially important for aptitude tests and for other tests used to assist in selection or placement)

12. Summarize results of statistical studies *relevant to your intended use of the test.*
- NOTE: Report criterion, data regarding number of cases and significant characteristics of validation group, and results of validity studies.^b

^a Many universities and colleges maintain files of standardized tests for use by students in measurement classes. If such a file is not available, or if the file does not include tests in a student's particular field of interest, the student in a measurement class is given the privilege of ordering directly from test publishers specimen sets of tests that he wishes to evaluate. Such requests, of course, must be countersigned by the professor, who will check the request for its appropriateness.

Table 5.1 (Continued)Summary of Data Needed in the Appraisal of a Standardized Test^a

OTHER EMPIRICAL EVIDENCE OF VALIDITY

13. Summarize other validation studies that help in the interpretation of test data *for intended use*.
 14. Your comments regarding statistical evidence of validity of test *for intended use*.
-

RELIABILITY (for total scores and for any subscores that will be interpreted)

15. Evidence of equivalence or internal consistency (that is, consistency of performance on specific content samples).^c
 16. Evidence of stability (consistency over time).^d
 17. Comments regarding adequacy of reliability *for intended purpose*.
-

NORMS

18. Types of converted scores.
 19. Availability of multiple norms for homogeneous subgroups (for example, by sex, age, occupation, curriculum, and the like).
 20. Adequacy of norming sample(s).
 21. Recency of norms (date of latest revision).
 22. Your comments regarding adequacy of norms *for intended use*.
-

PRACTICAL CONSIDERATIONS WITH RESPECT TO ADMINISTRATION AND USE

23. Complexity of administrative process for examiner and students.
 24. Time requirements.
 - Working time for students
 - Total administration time
 - Is more than one testing session required?
 25. Scoring.
 - Have adequate procedures been used to minimize scoring time?
 - If the scoring is not entirely objective, does manual provide adequate directions for scoring?
 - Are any special qualifications required for scoring the test?
 26. Aids to interpretation.
 - Can raw scores be easily translated into converted scores appropriate to your purpose?
 - Are special forms (such as profile sheets) provided to aid in the interpretation of results? Are these forms so designed as to help the user consider errors of measurement?
 - Does the manual provide sound and helpful aids to interpretation and valid suggestions for use?
 27. Cost of testing.
 - NOTE: Consider not only cost of booklets but whether such booklets can be re-used with consumable answer sheets. Consider also clerical time involved in scoring.
-

Table 5.1 (Continued)

Summary of Data Needed in the Appraisal of a Standardized Test^a

COMMENTS OF REVIEWERS (See test list in Appendix for references to reviews in Buos Yearbooks)

YOUR OVER-ALL EVALUATION OF THE TEST FOR INTENDED USE

^a In order to conserve space in printing, no space has been allowed for filling in data or comments. The student would merely incorporate the headings in his typed report.

^b In interpreting validation data, take the following factors into account: (1) the criterion variables used (evidence concerning their reliability and probable relationship to ultimate criterion); (2) time elapsed between administration of predictor test and obtaining of criterion scores; (3) evidence of possible criterion contamination (for example, test data being available to persons making criterion ratings); (4) characteristics of validation group (number of cases, *M* and *SD* of test and criterion scores); (5) whether the test was cross-validated, that is, validity coefficients were computed on a different group than the one used in the selection of test items; (6) whether data are given that would enable the user to judge the confidence with which he can estimate criterion data.

^c In interpreting data, note the number of cases, the method used, and the results. Also note whether the reliability coefficient was computed on groups that are about as homogeneous as the groups on which a test is typically used; (for example, a median reliability coefficient for several school-grade groups is desirable, rather than a spuriously high reliability coefficient, computed on all such groups combined).

^d In interpreting data, note procedures, characteristics of group used in stability study, and the results. In addition, note the time interval between the two administrations of the test.

In Table 5.1 is presented a form for summarizing data about specific tests. In this outline, we have reminded the student of specific points studied in Chapters 2-4 that are relevant to test selection. In each section of the form, the student is asked to keep in mind his intended use (or uses) of the test data.

The first major topics in the outline are concerned with validity. Validity for the intended use is the *sine qua non* of any test. An achievement test that does not test learnings relevant to the goals of instruction may give misleading results; an aptitude test that does not have predictive validity may lead to the wrong decisions.

Validity actually includes reliability as well as relevance. A test must have a fair degree of reliability, that is, must measure some attribute with fair consistency, in order to provide a dependable basis for any type of judgment. Reliability is a necessary, but not sufficient, condition for validity. Following validity and reliability, the outline considers norms and practical considerations with respect to administration and use. Each major section of the outline will be discussed in turn.

Content Validity

In studying the *content validity* of a test for one's own purposes, it is essential to study the test and manual carefully. Check the manual to see if it provides a classification of test items to help the user judge content validity for his own purposes. Determine how closely the distribution of items (by content area and type of objective) agrees with the proportional emphasis desired. Check the test itself to determine what percentage of items appears valid for local use, to ascertain whether items seem to be well-constructed, and to determine the percent of test items that test understanding, rather than just requiring the students to recognize memorized content.

In the selection of achievement tests, *face validity* is important, that is, the extent to which a test *appears* to measure relevant information and abilities. Especially when an achievement test is used as part of a final examination, it is essential that the students feel that the test is fair in that it emphasizes what they have studied. A teacher can judge face validity, and also the appropriateness of the difficulty level of the test, only by an actual examination of test items. In fact, it is a good idea for the teacher to take the test.

Concurrent and Predictive Validity

Validation data of these two types are especially important when tests are to be used in making decisions *about* students or in improving the informational basis for decisions *by* students. We are especially concerned with these types of validity when decisions among alternatives are being made.

Examples of institutional decisions *about students* include the *selection* of students for college admission or other programs admitting limited numbers and the *placement* of students in ability groups. For such uses, *local* validation data and expectancy tables are indispensable. However, one can select tests that seem promising for local use by searching the test manual (or a technical supplement) for correlations between test scores and the criterion data in which one is interested, obtained on groups of students similar to local groups.

A test will not have high concurrent or predictive validity for a specific purpose unless the difficulty level is appropriate. If one's chief purpose is to differentiate among low-achieving students so as to select those who should be assigned to remedial classes, it is best to select a test "peaked" at a low level of difficulty—that is, with a large number of items that would constitute a good test for that group. In a test with insufficient "test floor," many students who would obtain scores scattered over a considerable

range on an easier test will obtain zero or near-zero scores. Such a test is of no value in differentiating *among* low-achieving students; we can have little confidence in predictions based on their scores.

On the other hand, if one's chief interest is in selecting students for an accelerated class, the test should be "peaked" at a high level of difficulty, containing a large percentage of difficult items. On a test with only a few difficult items, or too low a "test ceiling," high-achieving students (whose scores would scatter over a considerable range on a more difficult test) pile up on the high end of the scale, as in the arithmetic test (Figure 2.4). Such a test is of little or no value in differentiating *among* high-achieving students.

If we want to do a good job of measuring individual differences throughout a wide range of ability, we must use a fairly long test with a wide spread in item difficulty; or we can divide our group of examinees, with one group taking a higher-level test and another group a lower-level test.⁹

Other Empirical Evidence of Validity

In this section of the outline, one would summarize other validation studies that help the user to interpret the meaning or significance of test scores. Construct validity studies help in identifying the many factors that contribute to individual differences in test scores. Such studies should help one to determine whether test scores are influenced unduly by student differences in vocabulary level, speed of working, and other factors unrelated to the trait measured. In interpreting test scores, it is helpful to have research findings on whether or not significant differences are found between average scores made by boys and girls, by students from differing socioeconomic backgrounds, or by students who have or have not had specialized instruction. Answers to many such questions aid the test user in interpreting what the test scores mean.

The type of validation studies needed varies with the type of test. For example, for personality and interest inventories, it is valuable to have

⁹ Because test purchasers have wanted to simplify testing programs and get "maximum returns" for their testing money and testing time, they have encouraged publishers of achievement tests to spread their items very thinly over a wide range of both content and difficulty. As a result, we may have so few items appropriate for the lowest grade level of a multigrade range that a chance score (obtained by marking responses at random) can bring students almost up to grade level. Greater awareness of this problem, and the use of two or more tests (of different ranges in difficulty) seem essential if meaningful scores are to be obtained for all students in the usual heterogeneous groups included in city-wide testing programs. The practice used in the STEP tests, of planning directions and time limits so that tests of two or more levels can be administered at the same time, constitutes a useful precedent. This problem is discussed further in Chapter 13.

research findings on the effectiveness of the test in predicting relevant behavior outside the test situation; for achievement tests, one seeks data on the extent to which scores obtained by multiple-choice questions agree with results obtained by free-response tests of equivalent content. For any tests that are to be used in making inferences regarding intraindividual differences, data regarding the homogeneity of subtests and intercorrelations between subtests should be examined.

The topic of construct validity is so complex that the student is referred to the last section of Chapter 4 for a review of all the types of evidence that would appropriately be included in this section of the outline.

In evaluating evidence on concurrent, predictive, and construct validity, one must examine tables of validity coefficients for each test, select those coefficients that are most relevant to the intended use, and then compare them. Making sound inferences from comparisons of validity coefficients, however, is not an easy undertaking. In general, the higher the validity coefficient, the better; however, many factors have to be taken into account in making comparisons:

1. Criteria differ

For example, when the scores of college freshmen on the *Davis Reading Test* were correlated with grades in English, the average r was approximately .50; when scores on this same test were correlated with the *STEP Reading Test*, the correlations were .76 for the reading level scores and .81 for the speed scores.¹⁰ The lower correlation with teachers' marks is due, in part, to the lower reliability of the criterion and, in part, to the fact that many factors other than reading ability affect grades in English. Some tests lend themselves to easy validation; for others, such as the personality inventories, adequate criteria are difficult to find. For some tests the only feasible criteria are ratings (which are admittedly subjective and unreliable). One cannot expect high validity coefficients for this type of test.

2. Groups differ

The groups on which the correlations are computed differ. Correlation coefficients tend to be lower for more homogeneous groups.¹¹

3. When we attempt to interpret validity coefficients in terms of their value for us—in making the judgments we wish to make—we must ask ourselves how much additional information the test gives. For example, if we are trying to

¹⁰ Frederick B. Davis and Charlotte C. Davis, *Manual, Davis Reading Test* (New York: The Psychological Corporation, 1962), pp. 22, 26.

¹¹ For a discussion of the relationship between the heterogeneity of groups and the relative size of correlation coefficients, the reader is referred to Chapter 3, page 94.

predict success in algebra, and we already have the results of a fairly recent intelligence test on the cumulative record, we may be more impressed with validity coefficients of .40 to .50 between an arithmetic test and algebra grades than validity coefficients for another test of general scholastic aptitude, which range in the .50-.60 range. The reason is that the arithmetic test provides new information.

Reliability

In evaluating a test, our requirements with respect to reliability depend on the type of comparisons we wish to make. Comparisons between groups make the lowest demands on reliability; comparisons between individuals require greater reliability. If we wish to measure gains for individuals, we are dealing with difference scores that include the error variance of both tests; hence for such comparisons, reliability needs to be very high.

When we wish to study intraindividual differences in diagnosis or guidance, high reliability of subtests and fairly low correlations between subtests are essential. If test scores are to be used as a basis for making profiles and studying intraindividual differences, the manual should provide information concerning the *reliability of differences*. Profiles should be designed so as to minimize the risk of the user's attaching significance to small differences that might be reversed in direction on retesting. An example is the profile form for the *Differential Aptitude Tests* (Figure 6.2), on which a one-inch difference in the height of two bars may be interpreted as representing a reliable difference. Another technique is used in the STEP profile (Figure 5.2).

Another factor that affects the degree of reliability required is the finality of the decision and the use of other data in decision-making. If we are excluding students or applicants on the basis of a single test, reliability of scores should be very high; if we are using the test as a preliminary screening device and will examine other test and non-test data for those students whose scores do not give a clear-cut prediction, lower reliability is acceptable.

If we are making decisions *about* people as in college admission, ability grouping, and the like, minimum reliability requirements would be higher than if we were selecting a test to be used *by* people in making their own decisions. When a person is making his own choices, we can remind him of the standard error of measurement, and we can encourage him to take other data into account. Under such circumstances, we can accept tests with somewhat lower reliability than would be usable for institutional decisions. The student is referred to Chapter 3 for further suggestions regarding the interpretation of reliability coefficients.

Name LAWRENCE ALBERT E.
 School MIDTOWN H.S. Grade or Class 11
 Age 16 Date of Testing Fall 1962-63
 Years Months Fall or Spring

Norms Used

☒ Publisher's ☒ Fall Grade or Class 11
☐ Local ☐ Spring Other _____
☐ Other _____

© Copyright 1957, All rights reserved

Cooperative Test Division Educational Testing Service—Princeton, N.J., • Los Angeles 27, Calif.

Here you can profile a student's percentile ranks on as many as six tests in the STEP series. In order for your comparisons between the areas to be valid, all tests included should be administered within a period of 4 or 5 months.

Recording. Directions for recording information and drawing percentile bands on the PROFILE form are included in each STEP MANUAL FOR INTERPRETING SCORES. Consult the manuals for the tests used.

Interpreting. To compare a student's performance on one of the tests in the STEP series with that of students in the norms group used, the unshaded parts of the column above and below the percentile band are of no use. For example, if the Listening percentile band is 24-36, you know that 24 per cent of students in the norms group score lower than this student and 64 per cent score higher. In other words, this student's Listening performance is below average with respect to the norms group.

To compare a student's standings on any two tests in the STEP series, the following rules apply:

1. If the percentile bands for any two tests overlap, there is no important difference between the student's standings on those two tests.
2. If the percentile bands for any two tests do not overlap, standing represented by the higher band is really better than standing represented by the lower band.

Examples: According to local norms, a student's percentile bands for three tests are

Mathematics (2A) 50-63
 Social Studies (2A) 60-71
 Science (2B) 41-52

Bands for Mathematics and Social Studies overlap; there is no important difference between the student's standings in these two areas. The same is true of Mathematics and Science. However, bands for Science and Social Studies do not overlap; the student's standing in Social Studies is higher than his standing in Science.

More detailed discussions of interpretations are contained in each STEP MANUAL FOR INTERPRETING SCORES.

Printed in U.S.A.
 D111252A

STEP STUDENT PROFILE

SEQUENTIAL TESTS OF EDUCATIONAL PROGRESS

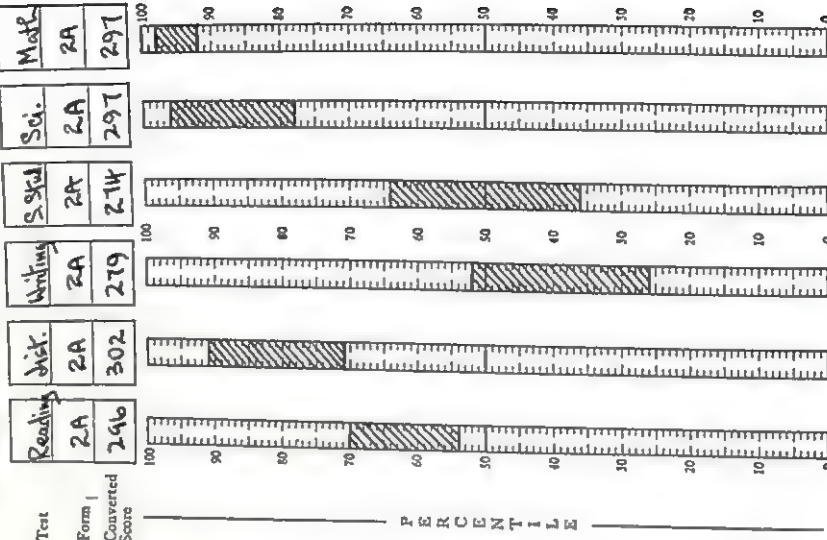


Fig. 5.2 Illustrative Profile for the STEP Tests for a Hypothetical Eleventh-grade Student

With respect to methods used in estimating reliability, it is ideal if two or more different methods have been used. By comparing the results obtained by different methods, we can estimate how much of the variance is due to the specificity of forms, how much to temporal variations in general characteristics of individuals, and how much to temporal variations in reactions to the specific test.¹²

If alternate forms of a test are to be used to measure growth, a coefficient of equivalence (obtained by the alternate-forms method) is essential; an internal consistency coefficient is not an adequate substitute. Internal consistency methods must not be computed for tests in which speed of working is an important factor in test scores.

Norms

Many standardized tests provide more than one type of converted score. For example, many achievement test batteries provide grade placement norms so that school averages can be compared with national norms in each subject area; and in addition, they provide percentile norms so that results for individuals can be interpreted in terms which students and parents can easily understand (provided that confusion with "percentage correct" scores is avoided).¹³

Normalized standard scores are becoming more widely used. Standard score units represent equal differences in raw scores throughout the score range. In other words, if three students each gained ten T-score points during the year, these gains would represent approximately the same gains in raw score, regardless of whether the student was initially in the low-scoring, average-scoring, or high-scoring group. This statement would not be true for *PR*'s; in fact *PR*'s are unsatisfactory as a measure of growth. Stanine scores combine the advantages of equality of units and ease of interpretation to lay persons.

If a test provides only one type of converted score, one must make sure that this type is adequate for the intended use. As we have mentioned, if a test has only percentile norms, interpretation of gains during a school year is difficult. A test that provides only grade placement norms may

¹² Cronbach presents an illustrative problem in which the four types of variance are estimated from data concerning coefficients of "equivalence," "stability," and "equivalence and stability." Lee J. Cronbach, *Essentials of Psychological Testing* (New York: Harper & Row, Publishers, Inc., 1960), pp. 136-138.

¹³ Ideally, percentile norms for school averages should also be provided. School averages do not vary as widely as scores for individuals. For example, if the average reading score for a school corresponded to a *PR* of 75 on the norms for individuals, that school would probably be achieving at a higher average level than 90-95 percent of the schools included in the norming sample.

give a misleading impression of the level of competency, and the relative competencies in subtests, for many students. However, if the test is satisfactory in other respects but lacks the type of norms desired, the development of local percentile and/or standard score norms might be advisable.

In appraising the adequacy of the norming samples used, one should consider not only number of cases but the procedures used to ensure that the norming sample is representative of the defined population. In norming achievement tests, the number of different communities used and their representativeness by geographic region, socioeconomic class, and other factors are important.

For some uses, the recency of norm revision is very important; for others it is less crucial. When one wishes to use "national norms" as a basis for evaluating achievement, a revision within the last eight to ten years is advisable, since publishers have used much more adequate procedures within the past decade for obtaining representative samplings of the general population.

At least three publishers have combined the norming of achievement batteries with the administration of their leading scholastic aptitude tests to the same students. Such dual standardization of tests of achievement and scholastic aptitude on the same norming samples helps the user to make sound inferences concerning individual differences between student achievement and aptitude.

Practical Considerations with Respect to Administration and Use

In addition to the major criteria discussed above, there are a number of minor criteria that should be considered in the selection of tests. The choice among two or three instruments that appear to be equally valid and reliable can be made on the basis of their usability or practicability—their cost, their mechanical make-up, the availability of equivalent forms, and the ease of administration and scoring.

COMPLEXITY OF ADMINISTRATIVE PROCESS A number of factors affect ease of administration—the clarity of instructions to examiner and subjects, the adequacy of sample exercises, and the requirement of close timing (involving the use of a stop watch). Examination of the test manual will provide the necessary information for judging ease of administration.

Clear directions are indispensable. Unless the instructions to the examiner and the directions to pupils are clear, the test is not sufficiently well standardized. Under these circumstances, one cannot be sure that he is giving the test under the same conditions as those for students in the norming sample. Adequate sample exercises are essential. If the type of items is familiar to students, one sample exercise may be sufficient; if the

type of item is unfamiliar or if the directions are complex, more than one example should be given.

Even before one orders specimen sets, one can usually determine from the publisher's catalog whether special qualifications are required for administering a test and/or interpreting the results. Several test publishers are now classifying their tests according to levels, as recommended by the American Psychological Association. These levels are:

LEVEL A Tests or aids that can adequately be administered, scored, and interpreted with the aid of the manual . . . (for example, achievement or proficiency tests).

LEVEL B Tests or aids that require some technical knowledge of test construction and use, and of supporting psychological and educational subjects such as statistics, individual differences, and psychology of adjustment, personnel psychology, and guidance (for example, aptitude tests, adjustment inventories with normal populations).

LEVEL C Tests and aids that require substantial understanding of testing and supporting psychological subjects, together with supervised experience in the use of such devices (for example, projective tests, individual mental tests).¹⁴

EASE AND OBJECTIVITY OF SCORING If the burden of scoring and computing converted scores is too great, the teacher has relatively less time for interpreting test results. For this reason, publishers are making every effort to reduce the scoring burden.

The use of test-scoring equipment, now available in many city and county school systems, requires special answer sheets, which are not ordinarily used below the fourth grade. Two leading types of scoring services are available: (1) the IBM Test Scoring Machines, which can be purchased or leased by school districts and (2) central scoring services that utilize special high-speed equipment, obtaining and recording several scores through a single "reading" (photoelectric scanning) of both sides of an answer sheet (which can accommodate almost a thousand test items).

One of the chief advantages of IBM scoring is that school districts can lease their own equipment, which can then be available for the scoring of teacher-made tests. One of the chief disadvantages is that answer sheets must be inspected to make sure that answer spaces have been fully blackened in and that all stray pencil marks and smudges have been removed. When sheets are well marked, 500 tests an hour can be scored with IBM machines.

¹⁴ "APA Code of Standards for Test Distribution," *American Psychologist*, vol. 5 (November 1950), pp. 620-626.

The photoelectric type of service is especially economical for achievement test batteries with several scores. Such a machine can score approximately 6000 answer sheets per hour, printing both raw and converted scores on a summary sheet. Answer sheets must be mailed to scoring centers.

An example of centralized photoelectric scoring is the National Guidance Testing Program of the Educational Testing Service, available for grades 4-14. A school district can obtain from one to nine scores for each student. The tests are scored and the program results are listed for each group. The basic service (scoring answer sheets, preparing list reports, permanent record slips, and individual report forms) is provided for a small per-pupil fee, *regardless of the number of tests the student takes* from the SCAT and STEP series. This example is given to illustrate how modern scoring and computing services make it possible to get many scores as cheaply as one or two. All leading test publishers now offer a similar type of scoring service for the tests they publish. Some publishers also offer hand-scoring services for tests given to pupils in the primary grades.

Some state universities, and some state and county departments of education, offer low-cost scoring services to school districts. A school that intends to do its own hand or machine scoring can estimate how time-consuming the scoring process is for each of several specific tests by comparing the fees charged by some nonprofit testing service, such as a state university that maintains scoring services for a wide variety of tests.¹⁵

Teachers have criticized machine-scoring methods because they do not indicate which of the students' answers are incorrect. Carbon-backed answer sheets have been developed to meet this objection. With one of these techniques, the Scoreze, developed by the California Test Bureau, the student marks his responses on a standard machine-scoring answer sheet, which can be detached and scored by machine if such equipment is available.¹⁶

MECHANICAL MAKE-UP The proper mechanical make-up of a test may be very important in its indirect effect on the validity of students' scores. The format should be attractive, and the size of type appropriate to the grade level. The quality of pictures and diagrams is very important.

¹⁵ An example of such a publication is *Unit on Evaluation, Test Scoring Service and Rental Service* (Champaign, Ill.: University of Illinois, 1955).

¹⁶ A somewhat different process has been developed by Harcourt, Brace & World, Inc. Students mark LDP answer sheets; the letters stand for the "Liquid Duplicating Process." The answer sheets are scored by overprinting the correct responses with a liquid duplicator. For some of the subtests, the master stencil overprints, at the same time, information pertaining to the skill or knowledge being measured by each test item. To date, this process is used only with the *Stanford Achievement Tests*, Elementary, Intermediate, and Advanced batteries.

NUMBER OF EQUIVALENT FORMS If the teacher wishes to measure growth in achievement by pre- and post-testing, it is necessary to administer two forms of a test, designed to be parallel in content and equal in difficulty. If a school staff plans to test intelligence or achievement at yearly intervals and compare the results for individuals and groups, it is important to select tests that have at least two equivalent forms. Several standardized tests have three to five such forms, equivalent in difficulty and designed to represent closely parallel samplings of skills and understandings.

AIDS TO INTERPRETATION Many publishers have developed excellent materials to aid teachers and counselors in the interpretation of test scores. The test authors are specialists in their own test and should be able to offer good advice concerning the ways in which the scores can best be interpreted and used.

Manuals for the *Sequential Tests of Educational Progress* (STEP) and for the *Evaluation and Adjustment Series* of high school tests¹⁷ provide item norms so that teachers can compare the achievement of their classes on each item with that of the norming sample. The *STEP Teacher's Guide* also provides suggestions for the discussion of test results with students. The manual for the *Metropolitan Achievement Test* is rich in sound suggestions for the use of results from this battery. The *California Achievement Tests* provide diagnostic analyses that assist teachers in obtaining diagnostic clues from student performance on small groups of similar items. These leads, however, should be checked with other sources of information, such as teacher-made tests or analysis of the student's work in the subject area.

Publishers of aptitude, interest, and other predictor tests have developed expectancy tables, multiple-group norms, student profiles containing aids to interpretation, and the like. For example, the publishers of the *Kuder Preference Record* (an interest inventory) have developed student reading lists on the different occupational fields, filmstrips and booklets to aid in the interpretation of results, and an "Occupational Counseling Review Set" for the counselor.

All these aids need to be evaluated, not just in terms of their number and attractiveness but in terms of whether they contribute to sound decision-making. Here, the opinions of test specialists, considered in the last section of this chapter, are of great value.

COST OF TEST SUPPLIES As a principal contemplates his dwindling supply budget, the relative cost of different tests may loom as an important factor in selection. However, one test may be costly at four cents per

¹⁷ These series are discussed in Chapter 13.

copy, whereas another one is cheap at fifteen cents. To waste teacher and student time on the administration and scoring of a cheap but inadequate test is poor economy. Since adequate standardization of tests is very expensive, the cost of tests involves more than the cost of paper and printing. Moreover, a more expensive test may include devices that reduce scoring time. The cost of scoring, which has already been discussed, may be a more significant factor than cost of materials.

For students in the upper elementary and the secondary grades, economy may often be effected by using separate answer sheets obtained from the publisher and reusing the more expensive test booklets again and again. Many publishers permit schools to lease tests; such an arrangement may provide an economical means of trying out a new test or of obtaining tests to be used in a special research study.

ILLUSTRATIVE USE OF THE SUMMARY FORM WITH A STANDARDIZED TEST

Before the student uses the suggested summary form in compiling and interpreting data regarding tests of his own choice, he should study the example of a completed form provided in Table 5.2. In this example, only the major headings of the outline have been used.

Table 5.2
Summary of Data Needed in the Appraisal of a Standardized Test
(Form filled in for the *Scholastic Aptitude Test*)

REFERENCE DATA

1. Title _____ *Scholastic Aptitude Test*
2. Names of major subtests: _____ *Verbal skills, mathematical skills*
3. Author(s) _____ *Staff, Educational Testing Service, with the advice of a committee of examiners in aptitude testing*
4. Publisher _____ *Educational Testing Service*
5. Range in grades: _____ *Applicants for admission to college*
6. No. of forms _____ *New form developed each year*

7. Purposes for which test is recommended by author(s):

"The specific job for which the *Scholastic Aptitude Test* was designed is to provide an indication of a student's ability to do college work. (It was not, and is not, expected to take over the whole job of assessing scholastic potential.) More precisely, the test is a measure of the level of development of the verbal and mathematical skills that are necessary to perform the academic tasks required in college."^a

Table 5.2 (Continued)

Summary of Data Needed in the Appraisal of a Standardized Test
(Form filled in for the *Scholastic Aptitude Test*)

CONTENT VALIDITY

This test is not intended to sample all aspects of intelligence, or even of scholastic potential. Many studies have been made over the years to improve the balance of content so that the test is not biased in favor of men or women, or in favor of students majoring in either the humanities or the mathematics-science curricula.

CONCURRENT AND PREDICTIVE VALIDITY

The criterion for validating the SAT tests has usually been grade-point average during the freshman year of college. However, in some studies, the four-year college average, graduation v.s. nongraduation from college, and grades in specific academic subjects have been used.

Studies of the relationship between SAT scores and college grades reveal a substantial relationship between SAT scores and subsequent academic performance in many different types of colleges, with verbal scores providing the better prediction in some curricula and mathematical scores in others. Separate prediction studies have been made for different types of colleges and different types of college curricula. Validity coefficients as high as .60 with freshman grade-point average are obtained under favorable conditions (that is, when college students follow a relatively uniform academic program, when college grades are based on extensive information about student performance, when grading standards are fairly consistent from one instructor to another, and when almost all students are working at a relatively high level of motivation).

Predictive Validity at Different Score Levels

Although research data seem to indicate as good predictive validity for low-scoring, average-scoring, and high-scoring students, a special study was made to see whether a high-level form of SAT was needed for the most able students. The high-level test devised did not show sufficiently improved validity to justify its use; but research on the best way to supplement the SAT with test data that will help highly selective colleges in selecting students from among high-ability applicants is continuing.

Predicting SAT Scores from SCAT Scores

Expectancy tables are available for predicting SAT scores from scores on the high school edition (SCAT), given in the 8th, 9th, 10th, and 11th grades.

OTHER EMPIRICAL EVIDENCE OF VALIDITY*Effect of Coaching on Test Scores*

An attempt has been made in successive revisions of the SAT to make the test as impervious to coaching as possible. Seven research studies (four made by the Educational Testing Service and three by independent investigators) agree that the average gain as the result of special coaching is less than 10 points, or 1/10 of a standard deviation.

Effect of Fatigue on Test Scores

Research studies have revealed no evidence that students perform less well on sections taken toward the end of a day of testing.

Table 5.2 (Continued)

Summary of Data Needed in the Appraisal of a Standardized Test
(Form filled in for the *Scholastic Aptitude Test*)

Effect of Anxiety on Test Scores

A half-hour version of the SAT was administered to 2000 students under the usual "anxious conditions" as part of the regular administration of tests for the College Entrance Examination Board. This same condensed version was administered to the same students under "relaxed conditions," the students being told that the study was being made for research purposes and scores would not be reported to the colleges. They were urged to do their best, however, since scores would be reported to their own high schools. Test anxiety seemed to have no effect on boys' scores but seemed to slightly increase girls' scores on the mathematics section. There was no difference in the concurrent validity of the tests, given under "anxious" as compared to "relaxed" conditions, that is, in the correlation of SAT with high school grades.

Fairness to Students from Different Cultural Backgrounds

Since the purpose of SAT is to predict the students' ability to do academic work in college, the question investigated was: "Do students from different or underprivileged backgrounds do better in college than one would expect from their SAT scores" or "Is the SAT really a fair measure of their ability to handle college work."

Research studies showed that, despite marked differences in background among students taking the test, there was no general tendency for students from different socioeconomic backgrounds to do any better or worse in college than the test scores predicted. Hence the College Entrance Examination Board concluded that "any cultural unfairness to students from less favored backgrounds who seek admission to college lies less in the *Scholastic Aptitude Test* than in the educational and environmental inequalities of our society."^b

RELIABILITY

Reliability coefficients have been computed by the Kuder-Richardson method on representative samples of students taking the test. Reliability coefficients in the most recent three-year summary ranged from .88 to .91 for both the verbal and mathematical tests. The standard error of a SAT score is approximately 30 points or 3/10 of a standard deviation. For example, if a student's "true score" were 500, the chances would be two out of three that his score on a single administration of SAT would lie between 470 and 530. This range, within which repeated scores for this student are likely to be obtained, represents a range of 10–15 percentile points.

The principle that a SAT score should be interpreted as a point within a range on the score scale, rather than a precise measurement, is emphasized in all interpretative materials sent to students, high schools, and colleges.

NORMS

College Board scores are reported on a scale based on a linear transformation of raw scores, with a mean of 500 and an *SD* of 100. The original scale was based on

Table 5.2 (Continued)

Summary of Data Needed in the Appraisal of a Standardized Test
(Form filled in for the Scholastic Aptitude Test)

students' score in the April 1941 test administration. As subsequent forms have been developed, each new form has been equated to the April 1941 form through an equating study, in which a sampling of new candidates answers a group of questions from previous forms so that adjustment can be made for any changes in the abilities of the groups from year to year.

Norms are available for many subgroups. For example, percentile ranks are available for all high school juniors, all high school seniors, high school students who choose to take SAT, all high school seniors going to college, and students enrolled at various types of colleges.

The College Entrance Examination Board provides distributions of SAT scores for admitted freshman students for all colleges with a large number of candidates. Results for these various norm groups are of considerable value to high school counselors in interpreting students' scores in comparison with their prospective peers in the different colleges they are considering. For example, a boy with a SAT verbal score of 500 would stand at the 85th percentile among high school seniors and at the 64th percentile among all students who enter college. In one of the moderately selective colleges, this student would rank above 47 percent of entering freshmen, but in another highly selective college, his score would exceed only 2 percent of entering freshmen.^c The sampling error involved in a single test of ability should be considered in making such interpretations.

PRACTICAL CONSIDERATIONS WITH RESPECT TO ADMINISTRATION AND USE

Time requirements: Three hours

Scoring: Machine-scored by central scoring service

Aids to Interpretation

An orientation booklet for students, entitled *A Description of the College Board Scholastic Aptitude Test*, including many sample questions

An interpretive leaflet to aid counselors in interpreting scores to students and parents, entitled *Your College Board Scores: Scholastic Aptitude Tests, Achievement Tests*

College Board Score Reports: A Guide for Counselors

College Board Score Reports: A Guide for Admissions Officers

Cost of testing: Paid by the student seeking college admission.

Source: Data excerpted, with the permission of the publisher, from "The Scholastic Aptitude Test," *Annual Report, 1961-1962* (Princeton, N. J.: Educational Testing Service, 1962), pp. 11-46; *The Scholastic Aptitude Test, 1926-1962, Test Development Report TDR-63-2* (Princeton, N. J.: Educational Testing Service, 1963).

^a "The Scholastic Aptitude Test," p. 11

^b *Ibid.*, p. 26

^c As listed in *College Board Score Reports: A Guide for Admissions Officers*. (Princeton, N. J.: College Entrance Examination Board, Educational Testing Service, 1962), p. 31.

The authors chose to summarize data regarding the SAT (*Scholastic Aptitude Test*), developed and administered by the College Entrance Examination Board. This choice was made because (1) the Board has recently issued a summary of their research on this instrument, which included a variety of validation studies; (2) the summarization of these data would be interesting to college students reading this textbook, because most of them have taken the test; (3) the choice of this test avoided the disagreeable alternative of reporting inadequacies in some commercially published test, which might be corrected within the next year or two. Under such circumstances, students would continue to study an erroneous test review, outdated by the revision of the test. Since many leading tests are in the process of revision, we would prefer to leave the critical review of published tests to the student, with the aid of his professor, the *Mental Measurement Yearbooks*, and other sources mentioned in the next section of this chapter.

SOURCES OF INFORMATION ABOUT PUBLISHED TESTS

In this textbook, we have deliberately minimized the discussion of specific tests. In Parts Two and Three, a few of the major tests will be discussed; but even here, little space will be given to sample profiles or to illustrative items from published tests. The authors believe that students should obtain and study actual tests and manuals. To assist them in finding tests of interest to them, a very comprehensive classified list of tests is presented in the Appendix.

One desirable outcome of a course in tests and measurements is the student's realization that he needs to consult expert opinion about tests in order to supplement his own indispensable study of a test and its manual. The student will find that his single best source for critical reviews of tests is the series of *Mental Measurements Yearbooks*, edited by Buros. The four most recent yearbooks, from 1940 through 1959, are listed in the Selected References at the end of the chapter.

When the student locates in the Appendix test entries in which he is interested, he will note in the right-hand column notations concerning all the test reviews that have appeared in the Buros Yearbooks. Reviews of a specific test may have appeared in more than one yearbook. Ordinarily, a new test is reviewed in the next yearbook that appears after its publication. However, Buros was not able to include in the early yearbooks (1938 and 1940) all of the tests then available. Hence, in succeeding yearbooks, he has included tests not previously reviewed, as well as additional reviews of widely used tests. As a result, the student can usually find two or more reviews of any test in which he is interested. In addition,

he may find in the yearbooks excerpts from reviews published in professional journals, as well as references to investigations in which the test has been studied.

The latest *Mental Measurements Yearbook* (1959) included only tests published through 1958; and for some of these, only bibliographical data could be included. For references to research studies on recently published tests, the reader should check appropriate chapters of the *Annual Review of Psychology* and those issues of the *Review of Educational Research* on psychological testing. The most recent issues on psychological testing were published in February 1959 and February 1962; further issues will appear at three-year intervals. The articles in the *Review* constitute brief, comprehensive surveys of research studies in each of the various fields of measurement during the preceding three-year period.

Since 1959, a test review section has been included in the *Personnel and Guidance Journal*; since 1954, *Personnel Psychology* has included a section called "Validity Information Exchange," which reports new data on the validity of tests used by personnel workers; in 1956 a similar information exchange on normative data was added. Validity studies are also summarized in two issues per year of *Educational and Psychological Measurement*.

A few books on measurement and evaluation in guidance have been published. Among the most recent are:

- Froehlich, Clifford P., and Kenneth B. Hoyt, *Guidance Testing*, 3d ed. (Chicago: Science Research Associates, Inc., 1959).
Goldman, Leo, *Using Tests in Counseling* (New York: Appleton-Century-Crofts, 1961).
Rothney, John W. M., and others, *Measurement for Guidance* (New York: Harper & Row, Publishers, Inc., 1959).
Super, Donald E., and John O. Crites, *Appraising Vocational Fitness*, 2d ed. (New York: Harper & Row, Publishers, Inc., 1962).

Additional books of value in the more specialized fields of aptitude, interest, attitude, and personality testing are included in the Selected References for Chapters 6 through 9 of this textbook.

Two books that would be of interest to teachers doing diagnostic and remedial work are the following:

- Blair, Glenn M., *Diagnostic and Remedial Teaching: A Guide to Practice in Elementary and Secondary Schools*, rev. ed. (New York: The Macmillan Company, 1956).
Bond, Guy L., and Eva Bond Wagner, *Teaching the Child to Read*, 3d ed. (New York: The Macmillan Company, 1960). Information is given in the Appendix concerning reading readiness tests, reading tests, and individual intelligence tests.

Books are available in several subject fields that focus on measurement and evaluation: Army, home economics; Hardaway and Maier, business education; Micheels and Karnes, industrial arts; and several in physical education, of which the most recent is by Clarke.¹⁸ Several recent yearbooks of professional organizations of teachers contain valuable chapters on measurement. The Educational Testing Service has reprinted such a chapter on evaluation in mathematics, with an annotated bibliography of published tests.¹⁹

SUMMARY STATEMENT

In this chapter we have illustrated how the principles presented in Chapters 2, 3, and 4 should be applied in the selection of tests for specific purposes; we have presented a test-evaluation form designed to assist the user in summarizing data relevant to test selection; and we have briefly reviewed sources of information that can be used to advantage in the selection process.

The principles developed in Chapters 2, 3, and 4 cannot be routinely applied. Experience is needed in their application to the appraisal of specific tests for specific purposes and groups. To aid the student in obtaining experience in appraising tests in his major field of study, a comprehensive, classified list of published tests is provided in the Appendix; references to reviews in *Buros' Mental Measurements Yearbooks* are included for each test for which such reviews were available (as of date of publication of this textbook).

In this chapter, assistance was also given in understanding the various types of tests available for use. The major characteristics which distinguish between standardized and teacher-made tests were considered. Tests were next classified according to the degree to which they involved indirect measurement or relevance to the criterion behaviors about which the test user wishes to make judgments. Several other bases for classification were clarified, namely: (1) group tests vs. individual tests, (2) pencil-and-paper tests vs. performance tests, and (3) speed vs. power tests.

When tests are classified on the basis of their content, the major distinction usually made is between tests of ability (in which the goal is to measure the individual's *maximum* performance) and inventories of personality, interest, and attitudes (in which the goal is to measure his *typical* performance). Although ability tests can be further classified into aptitude and achievement tests, these tests actually lie along a continuum with respect to their emphasis on verbal-educational factors.

¹⁸ Clara Brown Army, *Evaluation in Home Economics* (New York: Appleton-Century-Crofts, 1953); Mathilde Hardaway and Thomas Maier, *Tests and Measurements in Business Education*, 2d ed. (Cincinnati: South-Western Publishing Company, 1952); William J. Micheels and M. Ray Karnes, *Measuring Educational Achievement* (New York: McGraw-Hill Book Company, Inc., 1950); H. Harrison Clarke, *Application of Measurement to Health and Physical Education*, 3d ed. (Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1959).

¹⁹ Sheldon S. Myers, "Evaluation in Mathematics," *Twenty-sixth Yearbook*, The National Council of Teachers of Mathematics (Washington, D. C.: The Council, 1961). Reprint available from the Educational Testing Service.

SELECTED REFERENCES

- BUROS, OSCAR K., ed., *The 1940 Mental Measurements Yearbook*. Highland Park, N.J.: The Mental Measurements Yearbook, 1941.
- , ed., *The Third Mental Measurements Yearbook*. New Brunswick, N.J.: Rutgers University Press, 1949.
- , ed., *The Fourth Mental Measurements Yearbook*. Highland Park, N.J.: Gryphon Press, 1953.
- , ed., *The Fifth Mental Measurements Yearbook*. Highland Park, N.J.: The Gryphon Press, 1959.
- , ed., *Tests in Print*. Highland Park, N.J.: Gryphon Press, 1961.
- KATZ, MARTIN R., *Selecting an Achievement Test: Principles and Procedures*. Evaluation and Advisory Service Series No. 3. Princeton, N.J.: Educational Testing Service, 1958. Available on request.
- SUPER, DONALD E., AND JOHN O. CRITES, *Appraising Vocational Fitness*, rev. ed. New York: Harper & Row, Publishers, Inc., 1962, Chapter 3.
- WESMAN, ALEXANDER G., "Comparability vs. Equivalence of Test Scores," *Test Service Bulletin* No. 53. New York: The Psychological Corporation, 1958. Available on request.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. In what ways do standardized tests differ from teacher-made examinations?
2. With the aid of the summary form presented in this chapter, evaluate two or more standardized achievement tests.
3. Select a test in your subject field from those reviewed in Buros' *Mental Measurements Yearbooks*. Summarize and evaluate the Buros' reviews of the test you select.
4. Why do so many standardized tests involve indirect measurement of criterion behavior?
5. Examine two or three catalogs issued by test publishers. Do they restrict sale of their tests to persons qualified to administer them and interpret the results? Do they follow the APA code exactly?
6. What are the chief differences between ability tests and tests of typical performance?

PART TWO

The Study
of Individuals

Some people think of aptitudes as innate abilities. There is increasing awareness, however, that we inherit structures with potentialities for functional use, rather than abilities; and that the development of one's potentialities depends upon environmental factors. We can measure only the student's performance in his developed abilities, which represent the cumulated results of interaction between innate structures and environmental situations.

The problem of estimating the percentage of variance in general mental ability, which is attributable to hereditary vs. environmental factors, has appealed to many investigators.¹ However, the current trend is to devote greater attention to studies on the development of concepts and problem-solving abilities and to factors that seem to encourage independence, flexibility, and resourcefulness in children's problem-solving behavior.²

THE CONCEPTS OF APTITUDE AND ACHIEVEMENT

Aptitude is defined in the *Dictionary of Psychology* as "a condition or set of characteristics regarded as symptomatic of an individual's ability to acquire with training some (usually specified) knowledge, skill, or set of responses such as ability to speak a language, to produce music, and the like."³

¹ Cyril Burt, "The Inheritance of Mental Ability," *American Psychologist*, vol. 13 (January 1958), pp. 1-15; *Intelligence: Its Nature and Nurture*, 39th Yearbook (Chicago: National Society for the Study of Education, 1940).

² John McV. Hunt, *Intelligence and Experience* (New York: Harper & Row, Publishers, Inc., 1961).

³ H. C. Warren, *Dictionary of Psychology* (Boston: Houghton Mifflin Company, 1934), p. 18.

When we use tests as measures of aptitude, we are concerned with how well test scores predict the examinee's *future performance* in some activity. When we use tests as measures of achievement, we are concerned with the examinee's present level of performance, as a basis for judging his success in *past* learning activities.

The way in which achievement and aptitude tests overlap in content and the extent to which tests vary in their emphasis on specific learnings were illustrated in Figure 5.1. Aptitude tests tend to be limited to tasks that are (1) *equally unfamiliar* to all examinees (that is, novel situations) or (2) *equally familiar* (in the sense that all students have had equal opportunity to learn, regardless of their pattern of specific courses).

We tend to use the term "aptitude" in two different ways:

1. Sometimes we speak of a person's having considerable *aptitude for* a subject (such as reading, science, or mathematics) or for a vocation (such as law or teaching). In this sense of the word, "aptitude" connotes a *combination* of traits and abilities that result in the person's being well qualified for training in a subject, activity, or occupation.
2. At other times we use the term "aptitude" in a narrower, more scientific sense to mean a discrete, unitary ability, such as numerical ability or spatial aptitude, which has significance (in varying degrees) for a number of subjects, activities, and occupations.

When we develop a test to predict success in some activity, and we wish to use the test only for that purpose, we use the term "aptitude" in its first meaning, as a combination of abilities that characterize successful performers in that activity. We might devise a test composed of many items, each of which correlates with success in that activity. We need to make no attempt to organize those items into homogeneous groups, designed to measure distinct abilities. A test devised to measure aptitude, in the first sense of the word, would logically be a heterogeneous, omnibus type of test.⁴ Two individuals might earn equal scores on such an omnibus test on the basis of superiority in quite different abilities.

In predicting student success in school work during the elementary school years, we are usually satisfied with a scholastic aptitude test of the omnibus type. We need a test that provides a composite score on a combination of abilities related to success in the academic aspects of the school program. We do not need a profile of abilities since the abilities of young children have not yet become highly differentiated in terms of differing interests and the selection of differentiated courses and work experiences.

In high school counseling, however, we are usually attempting to predict student success in *many* different fields. Moreover, the subject or vocation in which we wish to predict success varies from student to student. Hence, for purposes of counseling high school students, we would prefer

⁴ The term "omnibus test" is defined and illustrated in the Glossary.

to measure aptitudes in the second sense of the word, as unitary abilities. We would prefer an aptitude test with subtests measuring discrete, factorially pure abilities. Then, these scores could be combined in various ways depending on their significance for different vocations or training programs.

As we review briefly the history of aptitude testing, the reader will note that "aptitude" has been used in both senses of the word. Binet, in his attempts to identify slow learners, developed a test that was intended to measure scholastic aptitude, or *aptitude for school learning*. No attempt was made to identify or measure discrete, unitary components of mental ability. The early group tests of mental ability, developed during and after World War I, were also omnibus tests, with the items having been selected in terms of their relationship to general level of scholarship, or success in training programs.

In World War II, however, the armed services were concerned not only with level of general mental ability but with aptitudes for many specific jobs. A limited pool of manpower had to be assigned to jobs in which each person was likely to do best. The vastness of the classification problem led to a realization that a custom-made test of aptitude for each type of job was not feasible. If aptitudes could be identified, in the scientific sense of the word, as discrete, unitary abilities, tests could be devised that were homogeneous in content and fairly independent of each other. Then, the prediction job could be reduced to manageable proportions. Recruits could be measured on a limited number of aptitudes; their scores on these tests could be combined and weighted in different ways so as to predict success in many different jobs. Hence, World War II saw the development of multiscore tests of mental ability, or aptitude test batteries. Such batteries were based on research in factor analysis and designed to measure aptitudes in the second or scientific sense of the word.

In this chapter, we will discuss first the tests of general mental ability. Then we will turn our attention to multiscore tests of different components of mental ability. In the multiscore tests, an attempt is made to test examinee performance in several fairly discrete aptitudes. Following the discussion of these two major approaches to aptitude testing, briefer consideration will be given to more specialized aptitude tests.

TESTS OF GENERAL MENTAL ABILITY OR SCHOLASTIC APTITUDE

Pioneer Work in the Testing of General Mental Ability

In 1904 the minister of public instruction in France became concerned about the high percentage of failure in the schools of Paris. He appointed

Alfred Binet to a commission to identify those pupils who were so mentally deficient as to require instruction in special classes. Hence the first attempts to measure intelligence were designed to measure aptitude for school work.

In collaboration with Theodore Simon, Binet developed in 1905 an individual intelligence test that was to be the prototype of many leading intelligence tests still in use and that earned him the title of "father of intelligence testing." The 1908 revision was an improvement over the 1905 scale. Binet assigned each of the 59 tests of this revision to an age level (from age 3 to age 13). He introduced for the first time the concept of "mental age." Further experimentation resulted in the 1911 revision, published in the year of Binet's death.

The most widely accepted revision of the Binet scale was the Stanford Revision, published by Lewis M. Terman in 1916. It was accompanied by an extensive manual, providing a standardized technique for administration and scoring.⁵ For more than 20 years the Stanford-Binet was the standard measure of intelligence, the criterion with which all other intelligence tests, group and individual, were compared.

In 1912 William Stern proposed the idea of computing the ratio of mental age to chronological age as a measure of rate of mental growth. He called this ratio a mental quotient. Terman adopted this concept, which has since gained universal acceptance, and applied the term "intelligence quotient," or "IQ," in the Stanford revision of the Binet.

In an attempt to measure the intelligence of deaf children, the *Pintner-Paterson Performance Scale* was developed in 1917. The scale included a series of 15 picture puzzles, form boards, and other tests of a nonverbal character. Among the performance tests that have since been developed, the *Arthur Point Scale of Performance Tests* is probably the most widely used. Performance tests not only are indispensable in the testing of deaf children but also serve as useful supplementary material in the testing of young children, as well as bilingual subjects and mental defectives of all ages.

Development of Group Tests of General Mental Ability

In 1917, the United States government faced the problem of training a large army as quickly as possible and of selecting from this large group the men who should be trained as officers. The American Psychological Association offered its services to the government, and a committee of psychologists prepared the first *group test* of intelligence, the *Army Alpha*, which was administered to nearly two million men.

⁵ Lewis M. Terman, *The Measurement of Intelligence* (Boston: Houghton Mifflin Company, 1916).

The testing of illiterate and non-English-speaking soldiers, however, was still a problem, for performance tests had to be administered individually. The development of the *Army Beta* marked another milestone in testing in that it was a group test that involved performance-type or nonverbal items. The directions for the *Army Beta* could be given by means of pantomime.

The use of the verbal *Army Alpha* and the nonverbal *Army Beta* demonstrated (1) the value of mental tests for revealing individual differences in mental ability among people of normal intelligence; (2) the fact that mental testing need not be a costly, individual procedure; and (3) the value of the tests in the practical classification of men. Within the short span of two or three years, group intelligence tests were accepted to an extent that would probably not have been attained in a decade or more of civilian use.

Later Developments in Individual Tests of General Mental Ability

During the 20-year period following its publication, the Stanford Revision of the Binet test was widely used in schools and clinics, as well as in educational research. In 1937, a revision of the *Stanford-Binet Scale* was published,⁶ with two equivalent forms, L and M. The 1937 tests were made less verbal at the lower levels, and the earlier emphasis on rote memory at the upper levels was corrected.

As a contribution to the measurement of adult intelligence, Wechsler published in 1939 the *Wechsler-Bellevue Scale for Adolescents and Adults*.⁷ The scale included six verbal tests and five performance tests. An attempt was made to include the types of tasks that would interest adults. Norms were provided for persons from ages 10 to 60, the scores for each person tested being compared with those of others in his age group. Thus, Wechsler redefined the intelligence quotient in terms of the relative rank of an individual in a group of persons of approximately the same age.

An outstanding innovation was the fact that three intelligence quotients could be obtained from the test—one from the verbal subtests, one from the performance subtests, and a third from the total scale. Wechsler also emphasized the diagnostic value of the pattern of 11 subtest scores; he developed equated scores for his subtests so that a profile of relative abilities and disabilities could be drawn for each examinee. Later research studies revealed, however, that differences between subtest scores had low

⁶ Lewis M. Terman and Maude A. Merrill, *Measuring Intelligence* (Boston: Houghton Mifflin Company, 1937).

⁷ David Wechsler, *The Measurement of Adult Intelligence* (Baltimore: The Williams and Wilkins Company, 1939).

reliability; and pattern analysis proved to have low validity. Hence, this test cannot be considered a multiscore test of mental abilities.

The Wechsler-Bellevue has now been superseded by two batteries, the *Wechsler Adult Intelligence Scale* (WAIS) for ages 16 and above, and the *Wechsler Intelligence Scale for Children* (WISC) for ages 5–15. The WISC was published in 1949, while the revised adult scale (WAIS) was published in 1955. In Table 6.1 the characteristics of the Wechsler tests are summarized, in comparison with those for the Stanford-Binet.

Table 6.1
Comparison of the Stanford-Binet Scales and
the Wechsler Intelligence Tests

Stanford-Binet ^a	Wechsler ^b
TEST CONTENT AND ORGANIZATION	
Terman selected a wide variety of test items that measured general mental ability but made no attempt to include a sufficient number of items of any one type to justify computation of subscores.	Wechsler selected items that represented different types of intellectual performance. His tests include a broader range of tasks than does the Stanford-Binet.
Tests highly weighted with verbal abilities.	Five performance tests included in both WAIS and WISC.
Items arranged in a spiral omnibus form, according to increasing degree of difficulty. Items organized by age levels.	Items organized by types into subtests, grouped into a verbal scale and a performance scale.
Specific tests vary from one age level to another. Some types of items (for example, vocabulary and memory for digits) are used over a wide range of age levels; others appear at only one or two age levels.	The same types of tasks are used at all age levels within WAIS and WISC.
Items of wide range of difficulty included. Children with a mental age as low as two or three can be adequately tested.	Insufficient content is included of right difficulty level for testing children with mental age below 7.
Illustrative items given in Table 6.2 Items were selected that appeared to measure mental ability and on which success was highly correlated with age. ^c Items were selected for	The verbal scale of WAIS includes: general information, general comprehension, arithmetical reasoning, similarities, digit span, and vocabulary.

Stanford-Binet^aWechsler^b

the 1960 edition on the basis of high correlation with success on total test.

VALIDITY

Since the Stanford-Binet was the only widely used individual intelligence test for many decades, it has been the criterion for determining the concurrent validity for many group tests.

The predictive validity of the test in predicting academic achievement in school has been well established, with r 's approximating .70 at the elementary school level and .60 at the high school level.

Factor analysis studies indicate that the test measures somewhat different abilities at different age levels, for example, reasoning factors that appeared at three other age levels did not appear in the age 11 tests.^f

May yield invalid scores for bilingual children and others whose experience with the English language is limited; performance tests should be used with these children to obtain supplementary data.

RELIABILITY

Equivalent-form reliability coefficients computed separately for each age level. Median reliability coefficient for ages 2-6, .88; for older children, .93. Especially reliable for examinees of low mental ability. Since Form L-M includes only items showing

In WISC, the digit span test is optional. The performance scale of WAIS includes: digit-symbol substitution, picture completion, block design, picture arrangement, and object assembly.

In WISC, the coding of a simple message is substituted for the digit-symbol test; a maze test is added as an optional test.

WISC full-scale IQ correlates highly with Stanford-Binet, usually showing r 's in the .80's. The verbal scale correlates more highly with the Stanford-Binet than does the performance scale.^d

The predictive validity of the Wechsler tests for academic criteria is somewhat reduced by the inclusion of the performance tests that are slightly less reliable and are less closely related to academic success.

Factor analysis indicates that the Wechsler tests are factorially complex and that subtests do not measure pure factors.^e

Performance scale of special value for examinees with a language handicap; observance of subjects at work on performance tasks may provide clues to a qualified psychologist concerning emotional disturbance or brain damage. Routine, objective analysis of patterns of subtest scores is of questionable validity.

Split-halves reliability coefficients obtained—.92 to .95 for full-scale, .88 to .96 for verbal scale, .86 to .90 for performance scale. Performance test probably most reliable performance test now available. Differences of less than 15 points between verbal

Table 6.1 (Continued)
Comparison of the Stanford-Binet Scales and
the Wechsler Intelligence Tests

Stanford-Binet ^a	Wechsler ^b
<p>high relationships to total score, reliability coefficients for 1960 edition should be higher. Reliability coefficients lower for younger children.⁵</p>	<p>and performance IQ's should not be taken seriously. Reliability of differences between subtests too low to justify analysis by any objective method. Only differences larger than 3 scaled-score units should be taken seriously.</p>
NORMS	
<p>Norms for 1937 edition based on testing a sample of 3184 subjects, ages 1½ through 18, carefully selected to be representative of white, native-born population. Seventeen communities in 11 states were included, and an effort was made to have the sample representative with respect to socioeconomic level.</p>	<p>Norming samples for WAIS representative of general population (with sampling at each age level stratified with respect to region, urban-rural residence, occupation, and education).</p>
<p>Norms for 1960 edition are based on the 1937 standardization, adjusted in terms of data collected during the 1950s.</p>	<p>Norming samples for WISC drawn from 85 communities in 11 states.</p>
<p>Child's MA computed by adding to his basal age the number of months earned by tasks successfully performed at higher levels.^b</p>	<p>MA's not used.</p>
<p>IQ's on 1960 edition are standard scores with a mean of 100 and an <i>SD</i> of 16.</p>	<p>Verbal, performance, and full-scale IQ's are standard scores with a mean of 100 and an <i>SD</i> of 15. Subtest scores are converted into scale scores with a mean of 10 and an <i>SD</i> of 3.</p>
<p>A given raw score (MA) yields the same IQ at all adult ages.</p>	<p>A given raw score yields different IQ's at different ages, depending on its standard-score equivalent in the norms for that age group; for example, a specific raw score would yield a higher IQ at age 60 than at 20, for it would rank relatively higher with respect to the older group.</p>

Stanford-Binet^aWechsler^b

For children, adolescents, and young adults, Stanford-Binet IQ's average approximately 7 points higher than Wechsler IQ's.

USABILITY

Content interesting to most examinees. Special training and supervised experience in administering and scoring tests is essential. Manual provides adequate basis for attaining objectivity in scoring.

Diagnostic scores on the examinee's performance on different types of items *cannot* be obtained; however, observations of child's reaction to standard test situations is of help in diagnosis.

Usual procedure of having examinee continue until he fails all tasks at a grade level may be seriously upsetting to some children.

Content interesting to most examinees. Special training and supervised experience in administering and scoring tests is essential; directions somewhat less complex than in Stanford-Binet. Manual provides adequate basis for attaining objectivity in scoring.

Verbal and performance IQ's can be obtained. Scores on subtests can provide diagnostic clues, to be verified or rejected on the basis of other data.

Performance test scores of examinees affected little by language handicap. Performance tests provide good opportunity to observe behavior of subjects who are emotionally disturbed by their difficulties in problem solving.

^a Tests included for each of 20 levels of ability, ranging from tasks suitable for the average child of age 2, through four levels developed for differentiating among average and superior adults.

^b WISC (Wechsler Intelligence Scale for Children) for ages 5 through 15 and WAIS (Wechsler Adult Intelligence Scales) for ages 16 and above. Wechsler's selection of types of items was based on his observations that some mental functions are more disturbed than others in mental patients, and that some functions show a greater deterioration with age than others.

^c Items were selected that had a "steep age gradient." For example, a test item was considered a good one for seven-year-olds if few six-year-olds, most eight-year-olds, and about half of the seven-year-olds got it right.

^d Judith L. Krugman and others, "Pupil Functioning on the Stanford-Binet and the Wechsler Intelligence Scale for Children," *Journal of Consulting Psychology*, vol. 15 (December 1951), pp. 475-483; W. H. Guertin and others, "Research with the Wechsler-Bellevue Intelligence Scales, 1950-1955" *Psychological Bulletin*, vol. 53 (May 1956), pp. 235-257.

^e P. C. Davis, "A Factor Analysis of the Wechsler-Bellevue Intelligence Scale, Form I, in a Matrix with Reference Variables," *American Psychologist*, vol. 7 (July 1952), pp. 296-297; B.

Table 6.1 (Continued)
Comparison of the Stanford-Binet Scales and
the Wechsler Intelligence Tests

Balinsky, "An Analysis of the Mental Factors of Various Age Groups from Nine to Sixty," *Genetic Psychology Monographs*, vol. 23 (February 1941), pp. 191-234.

^f L. V. Jones, "A Factor Analysis of the Stanford-Binet at Four Age Levels," *Psychometrika*, vol. 14 (December 1949), pp. 299-331.

^g Quinn McNemar, *The Revision of the Stanford-Binet Scale: An Analysis of the Standardization Data* (Boston: Houghton Mifflin Company, 1942).

^h The basal age for a child is the highest age level at which he got all items correct. For example, if a child gets all the items correct at ages 5 and 6, his basal age is 6. If he gets two-thirds of the items correct at age 7, one-third of the items at age 8, and none at age 9, his MA would be $6 + 8 \text{ mo.} + 4 \text{ mo.} = 7 \text{ years.}$

In 1960, a new edition of the Stanford-Binet was issued. The items of forms L and M that had shown the highest validity (in terms of correlation with achievement on the total test) were retained in a single form known as Form L-M. In the 1960 edition, a set of tests is provided for each of 20 ability levels, beginning with tests appropriate for the average two-year-old and extending through four levels designed to differentiate among average and superior adults. Illustrative items for each of five levels are given in Table 6.2. Norms were revised in terms of data cumulated during the 1950's.

Table 6.2
Illustrative Items at Different Age Levels of the Revised
Stanford-Binet Scales

TWO-YEAR LEVEL

Identifying six parts of the body on a large paper doll

"Show me the dolly's hair." Credit for age 2 is given if child correctly identifies three; credit for age $2\frac{1}{2}$ if he correctly identifies all six.

Picture vocabulary

The child is asked, "What is this?" "What do you call it?" as he is shown 18 cards containing pictures of common objects (credit for age 2 is given for identifying two objects; credit for age $2\frac{1}{2}$ if eight are correctly identified).

SIX-YEAR LEVEL

Vocabulary

The same graded list of 45 words is used for age 6 and above. Credit at the 6-year level is given for six correct definitions of such words as "tap," "orange," "envelope," and the like. The examiner says, "When I say a word, you tell me what it means. What is an orange?"

Table 6.2 (Continued)
Illustrative Items at Different Age Levels of the Revised
Stanford-Binet Scales

Mutilated pictures

The child is shown five pictures, each showing an object that has a missing part, such as a wagon with three wheels. He is asked, "What is gone in this picture?" or "What part is gone?" Four out of five must be correct for credit.

Number concepts

Twelve one-inch blocks are put in front of the child. He is asked to give the examiner different numbers of blocks, for example, "Give me three blocks. Put them here." Four correct answers out of five give credit at the 6-year level.

Maze tracing

Three mazes (with starting and finish points marked) are presented in succession. In each maze one route is longer than the other. The child is asked to trace the shortest route. Two correct out of three give credit.

TEN-YEAR LEVEL

Vocabulary

Credit is given at the 10-year level if 11 or more words are correctly defined.

Word naming

The child is asked to name as many words as he can in two minutes. Credit at the 10-year level is given for 28 words or more.

Repeating digits

Six digits are read at one-second intervals. The child is asked to repeat them in exactly the same order. Three series are presented. Credit is given if the child recalls at least one of these series correctly.

TWELVE-YEAR LEVEL

Vocabulary

Credit is given at the 12-year level if 15 or more words are defined correctly.

Verbal absurdities

Five statements containing absurdities are presented. In each case, the examiner says, "What is foolish about that?" For credit, four of the five absurdities must be correctly identified.

Repeating digits reversed

The examiner says, "I am going to say some numbers, and I want you to say them backwards." Three series of 5 digits each are presented. The child is given credit if at least one of these series is correctly repeated.

Abstract words

The examiner asks, "What do we mean by *courage*?" Credit is given at this level for three correct responses out of four.

Table 6.2 (Continued)
Illustrative Items at Different Age Levels of the Revised
Stanford-Binet Scales

AVERAGE-ADULT LEVEL

Vocabulary

Credit is given at the average-adult level if 20 or more words are correctly defined.

Differences between abstract words

The subject is asked to distinguish between pairs of associated words, for example, "poverty" and "misery." Credit is given if the subject correctly distinguishes between at least two of the three pairs presented.

Proverbs

The subject is asked to explain the meaning of three proverbs. Credit is given for at least two correct interpretations.

Ingenuity

Three novel problems are presented, for example, how would one obtain exactly 3 pints of water from a river if one has only a 7 pint container and a 4 pint container. Credit is given for the correct solution of two out of three problems.

Later Developments in Group Tests of General Mental Ability

Many of the early group intelligence tests contained such a large percentage of verbal items that children with language handicaps or reading disabilities made spuriously low scores. Hence, educators welcomed group intelligence tests that provided language and nonlanguage intelligence quotients for children. One of the first group intelligence tests to incorporate this feature was the *California Test of Mental Maturity*. Samples of language and nonlanguage items from this test are given in Figure 6.1. The Pintner, the Lorge-Thorndike series, and others now provide separate IQ's on verbal and nonverbal tests.

Poor readers tend to obtain successively lower intelligence quotients on typical tests of general mental ability given in the upper elementary and high school grades as these tests include more and more items requiring reading ability. The use of a nonverbal intelligence test is helpful in determining whether the lower intelligence quotients obtained in the higher grades can be attributed chiefly to reading disability.

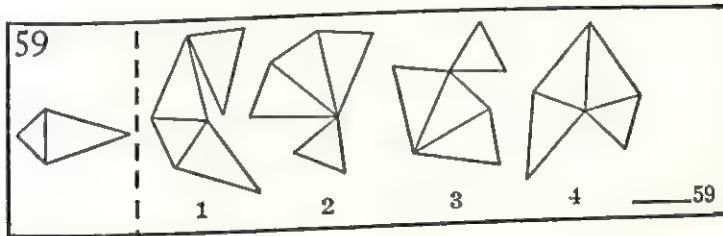
Actually, many so-called nonlanguage tests involve the use of considerable language in the giving of directions and in the child's use of verbal concepts in working with problems. However, reading ability is not required in these tests; hence they are considered to yield more valid estimates of learning ability for children who are poor readers. It is a mistake to assume, however, that the nonlanguage or nonverbal IQ represents the pupil's "true ability" or the level of performance he would achieve on

LANGUAGE ITEMS

147. How many $1\frac{1}{2}$ -cent stamps would you give in even exchange for 30 one-half-cent stamps?
 a. 10
 b. 15
 c. 20
 d. 45 _____147
215. A weighs less than B.
 B weighs less than C.
 Therefore
 1. B weighs more than C.
 2. A's weight equals B's and C's.
 3. A weighs less than C. _____215
220. W is between X and Y.
 X is between Y and Z.
 Therefore
 1. W is not between Y and Z.
 2. W is between X and Z.
 3. W is nearer to X than to Z. _____220

NONLANGUAGE ITEMS

In each row find the drawing that is a different view of the first drawing.



The first three pictures in each row are alike in some way. Decide how they are alike, and then find the one picture among the four to the right of the dotted line that is most like them.

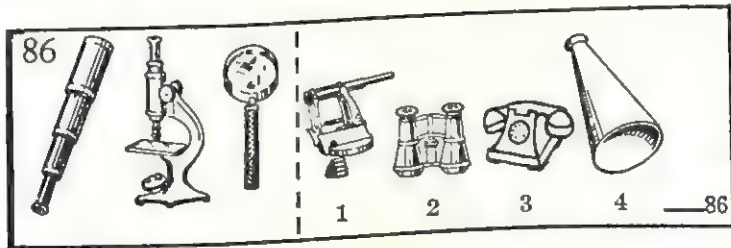


Fig. 6.1 Sample Items from the California Test of Mental Maturity

Reprinted with the permission of the California Test Bureau from Elizabeth T. Sullivan, Willis W. Clark, and Ernest W. Tiegs, *California Test of Mental Maturity*, Junior High, 1957 Edition. Monterey, Calif.: California Test Bureau, 1957.

verbal tests if his reading handicap were removed. Actually, nonlanguage tests measure different aspects of mental ability than the more verbal tests. They tend to have lower reliability and lower predictive validity for success in almost all school subjects, as well as in most of the vocations for which data have been studied. Low predictive validity especially characterizes nonlanguage tests that are limited to tasks involving perceptual skills and spatial aptitudes, rather than tasks that require reasoning with numbers or other symbols.

Individual vs. Group Tests of General Mental Ability

Children who have sensory handicaps,⁸ those who appear to be mentally retarded, and others who do not seem to perform adequately in a group testing situation should be given individual tests by trained examiners. In an individual test, the informal oral approach permits the examiner to establish rapport with the student, to stimulate maximal effort, and to observe and to evaluate his behavior.

The individual intelligence test offers several advantages: (1) greater variety of content, (2) opportunity to motivate the individual by means of encouragement and praise whenever necessary, (3) opportunity to adapt the tempo of testing to the personality of the individual, (4) absence of competition with others, and (5) greater opportunity to observe the individual's behavior under controlled conditions.

The disadvantages of the individual test lie in (1) the difficulty of administration, a trained examiner being required; and (2) the amount of time required for each administration, approximately one hour per subject. In city systems employing school psychologists, the use of individual intelligence tests is ordinarily limited to identifying the mentally handicapped, testing children with sensory handicaps, verifying the results of group intelligence tests for extreme deviates, and making diagnostic studies of children with special problems.

MULTISCORE TESTS OF MENTAL ABILITIES AND APTITUDE TEST BATTERIES

As a result of factor analysis studies by Thurstone and by many other factor analysts, considerable agreement has been reached concerning several components of mental ability. Table 6.3 lists 12 ability factors that have been confirmed by two or more factor-analysis studies.

⁸ For deaf children, the *Amoss-Ontario School Ability Examination* or the *Nebraska Test of Learning Aptitude* can be used. For those who are visually handicapped, the *Hayes Adaptations of the Binet and Wechsler-Bellevue Scales* are available. The *Arnold Adaptation of the Leiter International Performance Scale* and the *Columbia Mental Maturity Scale* can be used with the multiply-handicapped child.

Table 6.3
Ability Factors Confirmed by Two or More Factor Analysis Studies

VERBAL FACTORS

- V_c —Verbal comprehension—ability to understand words and written materials, such as in vocabulary tests (synonyms or antonyms, defining words), detecting absurdities in stories, sentence completion, reading comprehension.
- V_f —Word fluency—ability to produce words rapidly (as in rhyming, supplying synonyms for easy words, listing words in a category such as *foods*, or listing as many four-letter words as possible beginning with C).

REASONING FACTORS

- N —Numerical computation—speed of solving simple arithmetic computations.
- R_g —General reasoning—ability to invent solutions to problems, as in arithmetic reasoning problems.
- R_d —Deduction—ability to draw conclusions, as in logical syllogisms.
- R_e —Eduction of relationships—ability to see the relationship between two things or ideas and use this relationship to select other things or ideas, such as in verbal analogies.

MEMORY FACTORS

- M_r —Rote memory—ability to remember simple associations in which meaning is of little importance, for example, ability to study pairs of names and numbers, words and colors, and the like for a minute or two, and then show immediate recall of the pairs when only the names or words are given on the next page.
- M_m —Meaningful memory—ability to memorize meaningful material, such as sentences, lines of poetry, pairs of words that are meaningfully related, and the like.

SPATIAL FACTORS

- S_o —Spatial orientation—ability to detect quickly and accurately the spatial arrangement of objects with respect to one's own body, for example, detecting what maneuver an airplane is going through by examining a picture of the landscape from that vantage point. Spatial orientation seems to require an actual or imagined adjustment of one's own body.
- S_v —Spatial visualization—ability to imagine how an object would look if its spatial position were changed, for example, the examinee is shown a folded paper with several holes in it and is asked to choose one of four or five alternatives that shows how the unfolded paper would look.

PERCEPTUAL FACTORS

- P_s —Perceptual speed—ability to recognize perceptual details rapidly, especially similarities and differences between visual patterns, for example, checking pairs of letter groups or number groups that are identical (making no mark when they are different); choosing from several alternatives a geometrical form like the one first presented.
- P_c —Perceptual closure—the perception of objects from limited cues, that is, the mental "putting together" of a perceptual form, such as a word when only part of it is presented (as if partially erased or blurred).
-

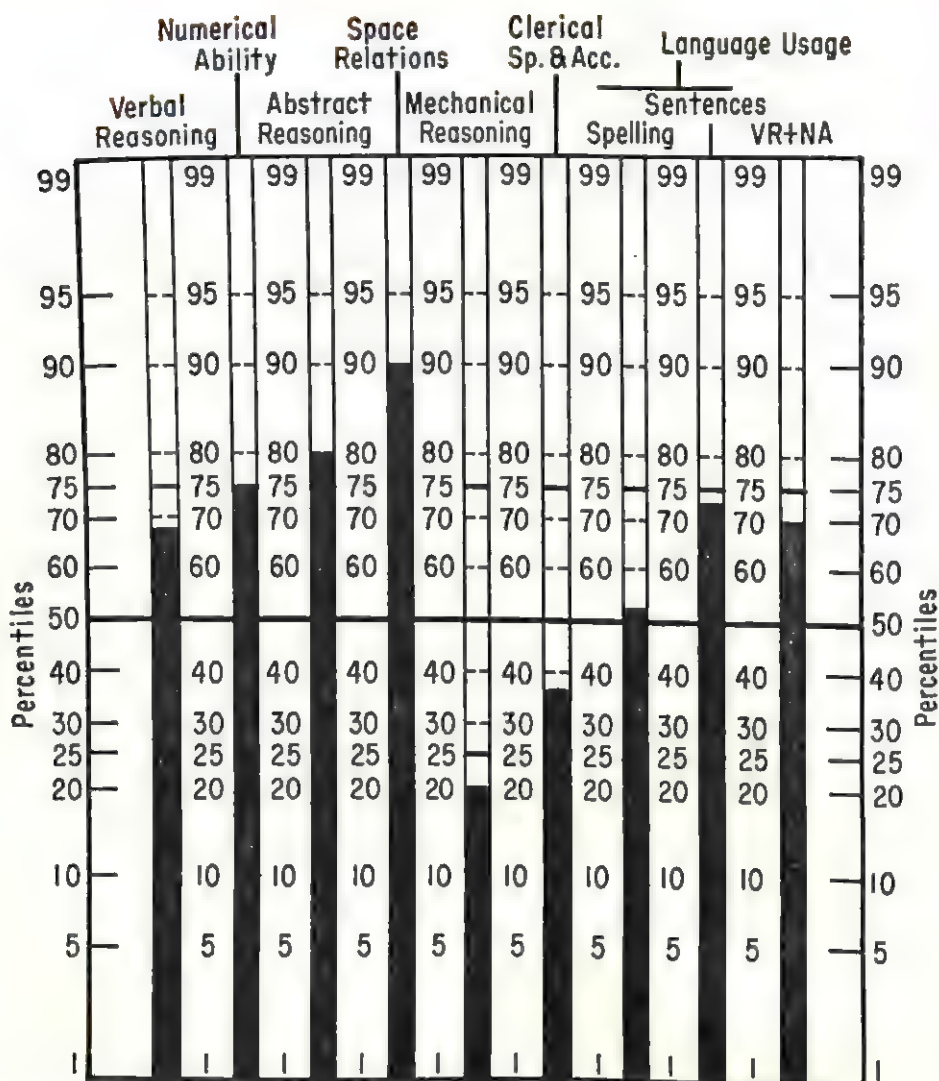


Fig. 6.2 Profile of Mary Dale's Percentile Ranks on the *Differential Aptitude Tests*

Adapted with the permission of the publisher from George K. Bennett and others, *Counseling from Profiles: A Casebook for the Differential Aptitude Tests* (New York: The Psychological Corporation, 1951), pp. 22-23.

NOTE: If the vertical difference between the heights of any two bars on the profile (before reduction) is one inch or more, the difference probably represents a real difference in the abilities measured (that is, the difference is significant at the 5-percent level). If the difference is between one-half and one inch, the student is instructed "to consider whether other things you know about yourself agree with it"; if the difference is less than one-half inch, the student is asked to disregard it as "probably not meaningful."

Mary Dale

PROBLEM

Mary sought help in her planning for higher education, having had to give up her ambition to study medicine.

TESTS

Differential Aptitude Tests, Grade 9.

An IQ of 111 was reported; test name and date of testing not given.

REPORT OF COUNSELING IN GRADE 12

Mary has a relatively good achievement record. She was undecided about her plans for college. Her initial goal had been to become a doctor. Because of the inability of her family to contribute financially to her college training, it was necessary for Mary to try for a scholarship. She failed in this competition.

Mary was assisted in understanding herself a little better through discussion of her *Differential Aptitude Test* results with her counselor. It is apparent from her tests that she does not have the superior aptitudes required of students who are awarded scholarships to prepare for medical training. She is now working and intends, after a year, to enter either a school for laboratory technicians or a medical secretarial school. In view of her persisting interests, her present plans seem more realistic. The test profile suggests adequate ability for either the technician course or the secretarial course—except that, for the former, a higher Mechanical Reasoning score might be desirable and, for the latter, her Clerical Speed and Accuracy and Spelling ratings are a bit low.

[EDITORS'] COMMENTS

The *Differential Aptitude Test* profile shows that Mary probably could enter a not-too-demanding college and do creditable work. Her present plans, however, seem to offer reasonable ways for her to satisfy her medical interests. Counseling should always take into account the motive behind the original choice of a goal; if Mary's desire to be a doctor reflected a concern for helping people, perhaps she would be happier as a social worker connected with a hospital or as a nurse. If her interests are primarily technical, then her present plans are probably wiser. Merely scaling the ambitions downward is not enough; positive suggestions for exploring alternates which are consistent with interests and abilities are just as necessary parts of a sound counseling process.

The foregoing comments refer to Mary's vocational plan. It should also be pointed out that a girl of Mary's abilities might find one, two, or four years of liberal arts or general education to be a valuable experience apart from vocational preparation. As a matter of fact, her vocational plan can be nicely integrated with a program in general education in either junior college or in a regular though not-too-demanding liberal arts college. Counseling problems are so often expressed in terms of vocational needs of young people that counselors sometimes forget the important values of education for citizenship and for developing mature cultural interests.

Tests of the various factors of mental ability can be especially valuable for use with high school students who seek assistance in choosing among various curricula and vocations. Instead of a single IQ, the tests provide a profile of students' scores on a test of several components of mental ability.⁹ An illustration of an aptitude test profile and its interpretation is given in Figure 6.2.

As more research data are made available, counselors can use the results of such multiscore or multifactor tests with increasing effectiveness as aids in counseling students. Until supporting research data are available, however, one cannot assume that high scores in verbal meaning predict success in linguistic pursuits or that high scores in space relations predict success in art and architecture.

When Thurstone completed his pioneer studies in the factor analysis of mental abilities, he developed a test battery designed to measure a number of primary factors that appeared to be sufficiently independent to justify their separate measurement. He designed the *Tests of Primary Mental Abilities*,¹⁰ which provide a profile of percentile ranks on six important factors of mental abilities—number, verbal meaning, space, word fluency, reasoning, and memory.

The PMA and other multiscore intelligence tests were developed largely on the basis of theoretical interest in the components of mental ability, while vocational aptitude test batteries were originally developed to meet practical problems of selection and classification. No clear-cut distinction, however, can be made between multiscore intelligence tests and aptitude test batteries. Both types of aptitude tests have been developed on the basis of research designed to identify and measure components of mental ability that (1) predict future achievement, (2) are relatively homogeneous, and (3) are fairly independent of one another. In each case, all the subtests are normed on the same standardization group, so that converted scores are comparable and a profile of the student's abilities can be drawn.

As predictive validity coefficients for success in vocations (or vocational training courses) are obtained for multiscore intelligence tests, and predictive validity data on success in various school subjects are obtained

⁹ In analyzing a student's profile of comparative achievements or abilities, it is important that the subtests be highly reliable and that they measure abilities that are fairly independent of one another (as shown by low intercorrelations). The problem of the statistical significance of differences between pairs of scores is discussed in Chapter 3.

¹⁰ The *Chicago Test of Primary Mental Abilities* (Chicago: Science Research Associates, Inc., 1941). This test was designed for ages 11–17. An abbreviated scale based on five of these factors has been developed, as well as tests of primary mental abilities for children of the elementary grades: *SRA Primary Mental Abilities*, ages 5–7, ages 7–11, ages 11–17. For further data on these tests and references to critical reviews, see Appendix A.

for vocational aptitude test batteries, the distinction between these two types of batteries becomes insignificant, despite their different historical origins. In fact, the publishers of the DAT, a vocational aptitude test, have suggested that the scores on tests VR and NA be combined as an index of scholastic aptitude. They have also developed a junior edition of the four subtests of the DAT that have the greatest predictive validity in educational decision-making and have published them as the APT (or Academic Promise) tests. Today the test user can simply choose the battery that has the greatest predictive validity for his own purposes without regard to the historical origins of the test.

The abilities measured by the PMA, a leading multiscore intelligence test, and, by the DAT, a leading vocational aptitude test battery, are compared below. It will be noted that in four of the subtests, the two

PMA ¹¹	DAT ¹²
(<i>Chicago Test of Primary Mental Abilities</i> , Single Booklet Edition)	(<i>Differential Aptitude Tests</i>)
Number	Numerical ability
Verbal meaning	Verbal reasoning
Space	Space relations
Word fluency	
Reasoning	Abstract reasoning
Memory	
	Mechanical reasoning
	Clerical speed and accuracy
	Language usage

batteries appear to measure similar abilities. The difference between these two batteries seems to be chiefly one of emphasis. The PMA includes tests of word fluency and memory, which probably have greater significance in predicting school achievement than vocational success, whereas the DAT subtests of mechanical reasoning, clerical speed and accuracy, and language usage are included primarily for their value in predicting success in a family or group of occupations (mechanical and clerical, respectively). Another difference is that the PMA attempts to measure relatively "pure" or uncorrelated mental abilities. Research studies, however, reveal that the subtests that presume to measure factors are still somewhat impure.¹³ Although the DAT battery claims to measure abilities that are

¹¹ Thelma Gwinn Thurstone and L. L. Thurstone, *Chicago Test of Primary Mental Abilities* (Chicago: Science Research Associates, Inc., 1941).

¹² G. K. Bennett, H. G. Seashore, and A. G. Wesman, *The Differential Aptitude Tests* (New York: The Psychological Corporation, 1947).

¹³ A. B. Crawford, and P. S. Burnham, *Forecasting College Achievement*. (New Haven, Conn.: Yale University Press, 1946).

relatively distinct, its emphasis—particularly in the last three tests—is on the prediction of success in a group of occupations rather than on the isolation of factors of mental ability.

These two tests also illustrate the difference between an aptitude test in which speed is a factor and one in which it is not. The PMA sets time limits that are so brief that a student's speed of work is a significant factor in determining his scores. In fact, Super and Crites contend that "speed plays too important a part in all the tests."¹⁴ With the exception of the subtest on clerical speed and accuracy, the tests of the DAT battery are power tests. The time limits are so liberal that almost all students are able to complete the tests within the time limits allowed. The relative desirability of speed and power tests of aptitude is best determined by their relative value for a specific prediction problem. For example, a speed test may be far superior for measuring aptitude for certain clerical functions (for example, rapid proofreading of names and numbers). On the other hand, a power test may be preferable for measuring more complex abilities, such as engineering aptitude or ability to learn foreign languages.

New forms of the DAT, incorporating minor revisions to facilitate administration and scoring, were published in 1963; the new standardization involved 45,000 students from 192 schools in 43 states. In Table 6.4 the DAT has been compared with the *General Aptitude Test Battery*, which has proved to be very valuable in the counseling of high school youth *who are not college-bound*.

Although the GATB was developed primarily for use by state employment services, these agencies are interested in helping high school seniors who are ready to enter the labor market. Where the GATB can be used effectively for this purpose, cooperative plans have often been developed on a local basis for the use of the battery in the schools.

Dvorak suggests nine steps for inaugurating a testing and counseling program that utilizes the GATB: (1) cooperative planning between representatives of the schools and the local employment service; (2) starting the program early in the last school year; (3) orientation of the students with respect to the program; (4) screening (in order to eliminate those students who are going on to college, who have made a final vocational choice, or who are not entering the labor market immediately); (5) administration of the tests (usually by employment-service personnel but sometimes by school guidance workers); (6) counseling interviews (by employment-service personnel, and sometimes also by school personnel); (7) transmission of records and information (with an interchange of test

¹⁴ Donald E. Super and John O. Crites, *Appraising Vocational Fitness by Means of Psychological Tests*, rev. ed. (New York: Harper & Row, Publishers, Inc., 1962), p. 137.

Table 6.4
Comparison of the DAT and GATB Aptitude Test Batteries

Differential Aptitude Tests	General Aptitude Test Battery
TEST CONTENT AND ORGANIZATION	
Eight tests, four measuring aptitudes in the strict sense of the term (VR—verbal reasoning, NA—numerical ability, SR—space relations, and perhaps AR—abstract reasoning); two that are factorially complex tests of aptitudes (MR—mechanical reasoning, CSA—clerical speed and accuracy) and two that are proficiency tests with predictive value (LU I and LU II—language usage: spelling and sentences).	Nine factor scores, obtained from 12 tests, as follows: G—general learning ability V—verbal aptitude N—numerical aptitude S—spatial aptitude P—form perception Q—clerical perception K—motor coordination F—finger dexterity M—manual dexterity
Measures selected variables known to be of value in educational and vocational counseling.	Measures most of the aptitudes that have been isolated and found to be occupationally significant.
Includes tests of mechanical comprehension and language, not included in GATB.	Includes tests of form perception, eye-hand coordination, motor speed, finger dexterity, and manual dexterity, not included in DAT.
TEST VALIDITY	
Tests of similar names evidently measure similar, but not identical, abilities, as evidenced by the following r 's between similar tests: verbal .72, space .72, numerical .62, and clerical .53. ^a	As basis for construction of battery, factor-analysis studies were conducted with 59 tests, administered in 9 overlapping batteries. On the basis of these studies, ten factors were identified and 15 tests chosen to measure them. In a later revision, the number of factors was reduced to nine, the number of tests to 12. Average intercorrelation between subtests is only .28.
Factor of general mental ability accounts for a substantial proportion of variance in all tests except CSA. However, tests are reasonably independent, the average intercorrelation between subtests being .38.	More work-oriented, developed by Occupational Analysis Division, United States Employment Service, primarily for use in vocational counseling of applicants for employment.
More school-oriented, designed primarily for use in high school counseling.	

Table 6.4 (Continued)
Comparison of the DAT and GATB Aptitude Test Batteries

Differential Aptitude Tests	General Aptitude Test Battery
Considerable data available concerning predictive validity of tests for grades in high school and college subjects. Has moderate differential validity for subject fields. Very little data available on concurrent or predictive validity for success in specific vocations or vocational training programs.	"Most adequately standardized and validated battery now available for vocational counseling and placement of inexperienced young persons and adults." ^b Considerable validation data (for 145 jobs and 23 occupational families). However, some studies provide only concurrent validity data. Typical validity coefficients average .50.
VR and NA score usable as measure of general scholastic aptitude.	G score usable as measure of general scholastic aptitude.
Except for CSA, all tests are power tests, with generous time limits.	Tests are highly speeded.
TEST RELIABILITY	
Tests long enough to give fairly reliable results. Average reliability coefficient .88 (Split-halves method used except for speeded CSA).	Tests shorter and somewhat less reliable, especially tests of perception, coordination, and dexterity (for which reliability coefficients range from .65 to .79).
High intercorrelations between certain tests reduces reliability of differences.	Low intercorrelations between most tests result in lower standard errors for difference scores.
Studies regarding stability of difference scores over a three-year period indicate that differences among CSA, MR, and the over-all level of the verbal-language-numerical tests are stable enough to be interpreted seriously.	Retest reliabilities over a three-year period (9th to 12th grade) almost as high as those for a three-month period.
Norms for each grade and sex (large, representative norming samples; in all, more than 45,000 students in 192 schools in 43 states involved in 1962 norming).	Multiple norms on many occupational groups. Number of cases in each occupational group is generally 50-200. Profiles for occupational families based on norming samples of 60-900. Critical scores on each of the three most crucial factors for each of 23 occupational families.
Percentile norms.	Standard scores with mean of 100 and SD of 20.

Differential Aptitude Tests	General Aptitude Test Battery
Norms by grade level indicate changes in aptitude with increasing maturity and experience.	Conversion tables make it possible for counselors to estimate individual's probable adult status on any test from his score as a high school student.
USABILITY	
All tests machine-scorable paper-and-pencil tests.	Eight of 12 tests machine-scorable; all quickly scored. Two tests involve use of simple apparatus, but can be administered in groups.
Alternate forms available for all tests.	Alternate forms available for first seven tests.
Any or all tests available for administration and scoring by any school district. All eight tests available in new two-booklet edition, used with only two answer sheets.	Tests given only through State Employment Service to high school students who plan to seek employment, rather than attend college. Cooperative plan makes results available both to high school counselor and employment service.
Spanish edition available.	Versions for use in at least 27 foreign countries have been prepared.

Source: Donald E. Super and John O. Crites, *Appraising Vocational Fitness by Means of Psychological Tests*, rev. ed. (New York: Harper & Row, Publishers, Inc., 1962), pp. 328-349; American Personnel and Guidance Association, *The Use of Multifactor Tests in Guidance* (Washington, D.C.: The Association, 1957); *Manual for the Differential Aptitude Tests*, 3d ed. (New York: The Psychological Corporation, 1957); and *Guide to the Use of the General Aptitude Test Battery* (Washington, D.C.: Government Printing Office, 1958).

^a *Guide to the Use of the General Aptitude Test Battery* (Washington, D. C.: Government Printing Office, 1958), p. L-2.

^b Super and Crites, *op. cit.*, p. 338.

data, personnel-record data, and the like between school and employment agency) and, ideally, case conferences on a number of students; (8) training of school personnel by state employment-service staff members in the use of the GATB and the various types of information available on the occupational and labor market and training of employment-service staff members by school personnel in the interpretation of previous test scores and other information on the school cumulative records; and (9) follow-up studies on the vocational adjustment of students tested, devel-

oped cooperatively by the school and employment-service personnel.¹⁵

Although these general policies have been outlined at the national level, it is recommended that any schools desiring to use the GATB consult the nearest office of their state employment service regarding the specific conditions under which the tests may be used.¹⁶

We shall not attempt to describe other aptitude test batteries. However, the subtests for each battery are listed in Appendix A. In making the difficult but highly important choice among aptitude test batteries, the teacher should consult the professional literature and revised test manuals for reports on research studies as well as reviews in the latest *Buros Yearbook*.

TESTS OF SPECIAL APTITUDES

Before aptitude test batteries were developed, the high school counselor who wished to do vocational aptitude testing had no choice but to administer a number of tests of single aptitudes (frequently called *special aptitude tests*). Outstanding examples of such special aptitude tests are the *Minnesota Clerical Test*, the *Revised Minnesota Paper Formboard*, the *Bennett Test of Mechanical Comprehension*, the *Meier Art Judgment Test*, and the *Seashore Measures of Musical Talents*. Each of these tests is one of the best in its field.

Although each special aptitude test may provide percentile norms, these percentiles are not based on the same norming population. Hence, unless local norms have been established, there is no basis on which the student's profile of relative abilities on several special aptitude tests may be drawn. On an aptitude test battery, however, subtest percentile ranks are comparable, since they have all been normed on the same students.

Special aptitude tests, however, may have the advantage of having been normed on several occupational groups or on groups of students enrolled in special curricula. Therefore, a student's score on, say, the clerical aptitude test can be compared with the scores of persons employed in routine clerical work, with those of bookkeepers, and the like. Tests of single aptitudes serve certain other functions, for example, to round out the picture of vocational strengths and weaknesses for individual students or to assist members of special departments, such as art and music, in

¹⁵ Beatrice Dvorak, "Proposal for Organizing a Multi-Factor Testing Program for Vocational Counseling," *Conference on Using Multi-Factor Aptitude Tests in Educational and Vocational Counseling and Prediction* (Berkeley, Calif.: University of California, Field Service Center, 1953), pp. 43-46.

¹⁶ A. W. Motley, Assistant Director, United States Employment Service, letter to the authors, April 29, 1955.

selecting and guiding students within their respective fields. No subtests on aptitudes in the arts are included in aptitude test batteries.

Performance Tests vs. Paper-and-Pencil Tests

Although aptitude test batteries are almost exclusively of the paper-and-pencil type, special aptitude tests may be of the performance type. Each has its special advantages and limitations.

In the strict sense of the word, performance tests require manipulative skills and involve the actual use of apparatus or materials. As such, they may be more concrete and meaningful to students and employees than pencil-and-paper tests, for the problems included in such tests closely resemble those involved in employment or training situations.

Performance tests have greater appeal, and therefore probably greater validity, for the less academic students. Paper-and-pencil tests are relatively abstract and require a higher degree of inductive reasoning on the part of the subject. Obviously, paper-and-pencil tests are less expensive than performance tests, for they can usually be administered in groups and can be scored quickly; consequently, they may be the only feasible choice in a school testing program. However, measurement of such abilities as manual dexterity, filing ability, and the like by means of paper-and-pencil tests is necessarily *indirect*, and the validity of such tests must be established through statistical studies of their ability to predict successful performance.

Figure 6.3 and the accompanying Table 6.5 offer a direct comparison between a performance test (the *Minnesota Spatial Relations Test*) and its paper-and-pencil "equivalent" (the *Minnesota Paper Form Board*). Careful study of these data indicates that the two tests are by no means identical. The performance test may be more useful in determining whether individuals desiring to enter certain semiskilled or skilled occupations have the required capacity for spatial visualization or manual dexterity. The paper-and-pencil test can be used for testing large numbers of students who wish to appraise their spatial judgment as one factor to be considered in their vocational planning.

Tests of Manual Dexterity

If the *General Aptitude Test Battery* (GATB) is given to high school juniors or seniors intending to enter employment, the performance tests in this battery will provide evidence concerning the student's motor coordination and his dexterity in finger and hand movements. If the GATB is not given, however, a counselor may wish to administer a test of manual

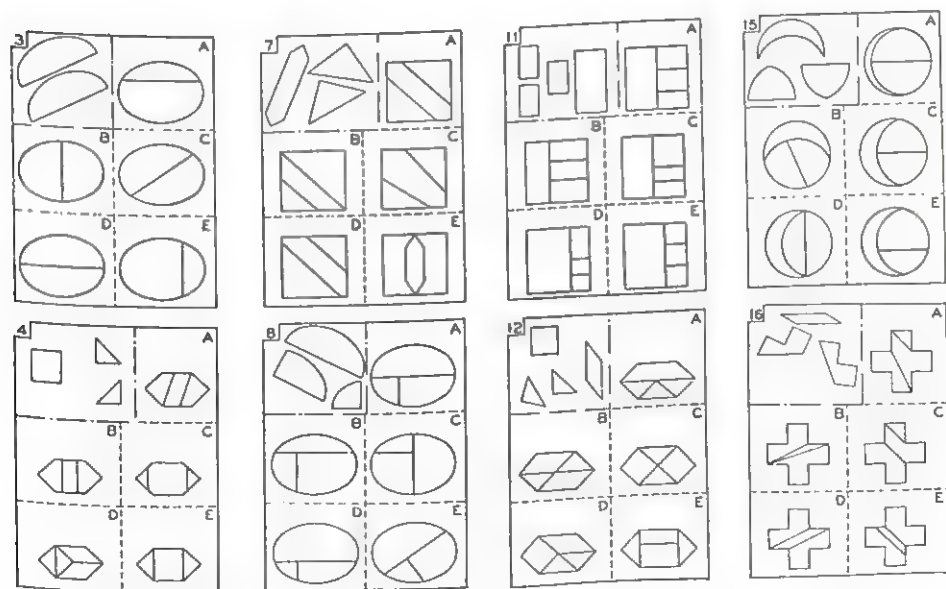
Table 6.5
Comparison of a Performance Test and a Pencil-and-Paper Test
in the Same Area

Minnesota Spatial Relations Test	Minnesota Paper Form Board
ADMINISTRATION	
Administered individually.	May be administered in groups.
SCORING	
Number of seconds required to complete last three form boards.	Number of correct responses in 20 minutes. May be machine scored.
CONTENT	
Four form boards, three feet long by one foot wide. Shapes include crescents, squares, and odd-shaped geometrical forms.	Sixty-four items. For each item, the "stem" consists of two to five disarranged parts of a geometric figure. The student chooses one of five responses (assembled geometric figures). The problem in each case is to select the figure that can be assembled from the parts. To make the appropriate figure, it may be necessary merely to push the parts together, or to turn them around or even over. All matching of shapes must be done <i>mentally</i> ; no trial-and-error work is possible.
AVERAGE TIME	
20 to 25 minutes.	Twenty minutes after practice problems.
FACTORS MEASURED	
Ability to visualize and judge spatial relations; ability to perceive spatial differences; reasoning.	Ability to visualize and judge spatial relations; perceptual ability; inductive reasoning. The reasoning factor is more heavily weighted than in performance test. Scores are mixed speed-and-level scores, with level of difficulty playing a lesser part.

Source: Data summarized from Donald E. Super and John O. Crites, *Appraising Vocational Fitness by Means of Psychological Tests*, rev. ed. (New York: Harper & Row, Publishers, Inc., 1962), pp. 281-300.



Fig. 6.3 A Performance Test (above, *Minnesota Spatial Relations Test*) and a Paper-and-Pencil Test (below, *Minnesota Paper Form Board Test*) of Spatial Relations



dexterity to a student for whom such information seems significant. For example, a student who is interested in dentistry and has all the academic qualifications might wonder if he has sufficient manual dexterity for such highly skilled work. Since such tests must usually be administered individually by someone who is specially trained, manual dexterity tests should ordinarily be reserved for the study of students with special problems in vocational counseling. Low scores must be interpreted in light of research concerning the trainability of certain skills.

Super and Crites have reviewed a number of validation and norming studies on the *Minnesota Rate of Manipulation Test*—a test of gross arm-hand dexterity.¹⁷ On the basis of their review, they doubt whether tests of gross manual dexterity have any value in the counseling of high school students. Their value lies chiefly in the selection or preemployment counseling of persons seeking such jobs as packing or large-part assembly.

Tests of finer dexterity, such as the *O'Connor Finger and Tweezer Dexterity Tests*,¹⁸ may be of value in counseling high school students desiring work in assembling small parts or in professions such as dentistry, which require fine manual dexterity.

The *Purdue Pegboard* measures both arm-hand dexterity (of a finer type than the Minnesota test) and finger dexterity (in a more realistic situation than the O'Connor tests). The fact that the student's ability to coordinate the action of both hands and to eliminate nonessential operations affects his scores may make this a more valid instrument for use in predicting success in certain occupations. Although high school norms are not now available, the early maturation of manual dexterity may make the adult norms applicable.

Tests of Clerical Aptitudes

Since several aptitude test batteries include tests of clerical aptitude, schools make limited use of special aptitude tests in this field.

Two widely used tests illustrate different approaches to the measurement of clerical aptitude. The *Minnesota Clerical Test* is a highly speeded, homogeneous test of clerical speed and accuracy, similar to those included in many aptitude test batteries. The *General Clerical Test* is a much more comprehensive test, measuring three types of ability important in office work: clerical speed and accuracy, numerical ability, and verbal facility. Obviously, the critical scores required in these three tests would vary considerably for different types of office jobs.

¹⁷ Super and Crites, *op. cit.*, pp. 182–217.

¹⁸ Johnson O'Connor, *O'Connor Finger and Tweezer Dexterity Tests* (Chicago: C. H. Stoelting, 1928); reviewed in Super and Crites, *op. cit.*, pp. 200–213.

APTITUDE TESTS IN ART Of special interest are the aptitude tests in the fine arts, because aptitude test batteries do not attempt to measure the abilities requisite for high-level performance in these fields.

The *Meier Art Tests: I Art Judgment*,¹⁹ designed for use in grades 7-12, require the student to select the more artistic picture in each of one hundred pairs of pictures. The test may be administered individually or to a group; the scoring is simple and objective. The test seems to measure a refined type of aesthetic judgment that matures during the secondary school and adult years.²⁰

A new *Aesthetic Perception* test has been completed by Meier and his associates; the final steps in the item validation and standardization of this test were completed in 1963. According to Meier, the new test "differs from test I . . . in the degree of penetration involved: the subject must evaluate four versions of each art product in order to rank them in order of aesthetic character."²¹

The Graves' *Design Judgment Test*²² has been designed to avoid the use of representational art. All of the designs are abstract. Each exercise of this test consists of two or three designs, one of which is organized according to the fundamental principles of art structure, the other design or designs violating one or more such principles. Norms are available for high school and college students. The test may be administered in 20 to 30 minutes and may be machine scored.

Aptitude Tests in Music

One of the earliest tests of special aptitude was the *Seashore Measures of Musical Talent*, developed in 1919. After the tests had been used for 20 years and extensive research on their use had cumulated, a revised edition was issued in 1939. The tests are now available on a single long-playing record and are adapted for machine scoring.²³ The following six elements are measured: (1) pitch, (2) loudness, (3) time, (4) timbre, (5) rhythm, and (6) tonal memory. In the first test (sense of pitch), for example, a series of paired sounds is played. The student is required to indicate for each pair whether the second sound is higher or lower in pitch than the first. Norms are available for grades 5 and 6, grades 7 and 8,

¹⁹ Norman C. Meier, *Meier Art Tests: I, Art Judgment* (Iowa City, Iowa: Bureau of Educational Research and Service, University of Iowa, 1940).

²⁰ Super and Crites, *op. cit.*, p. 309.

²¹ Letter to the author from Dr. Norman C. Meier, May 11, 1963.

²² Maitland Graves, *Design Judgment Test* (New York: The Psychological Corporation, 1948).

²³ Carl E. Seashore, Don Lewis, and Joseph C. Saetviet, *Seashore Measures of Music Talent*, rev. ed. (New York: The Psychological Corporation, 1960).

and adults. Research studies indicate that special training in music does not influence test scores.

The *Musical Aptitude Test*, developed by Whistler and Thorpe,²⁴ consists of short exercises played on a piano. The test includes subtests in (1) rhythm recognition, (2) pitch recognition, (3) melody recognition, (4) pitch discrimination, and (5) advanced rhythm recognition. Students' responses are recorded on answer sheets and can be machine scored. Percentile norms for grades 4 and above are provided.

The *Drake Musical Aptitude Test*,²⁵ measures two significant music aptitudes—musical memory and rhythm. Two equivalent forms of the test are available on a single long-playing record. The test can be administered to the least talented child and yet is difficult enough for the gifted adult. Results do not seem to be affected by musical training.

The *Wing Standardized Tests of Musical Intelligence*,²⁶ developed in England, includes seven tests that cover chord analysis, pitch change, memory, rhythmic accent, harmony, intensity, and phrasing. The first three parts require students to make sensory discriminations, but at a more complex level than the Seashore tests; while the other four require students to compare the aesthetic merits of pairs of selections of piano music. Norms are provided on total scores. The tests have a sufficiently high ceiling to differentiate among talented individuals. Validity studies with small groups have resulted in correlations of .60-.70 with teachers' ratings of musical ability.

PROGNOSTIC TESTS

The last type of aptitude tests we will consider illustrates the use of the term "aptitude" in its first sense, that is, aptitude *for* learning in some *subject field*. As our initial discussion suggested, these tests may be omnibus tests. They may include subtests on abilities considered necessary in the subject field, but no attempt has usually been made to have these subtests measure aptitudes in the sense of discrete, unitary abilities.

READING-READINESS TESTS Perhaps the most widely used tests in this group are the reading-readiness tests. Typically, they measure the pupil's knowledge of oral vocabulary, his ability to follow oral directions, to match

²⁴ Harvey S. Whistler and Louis P. Thorpe, *Musical Aptitude Test, Series A* (Monterey, Calif.: California Test Bureau, 1950).

²⁵ Raleigh M. Drake, *Drake Musical Aptitude Tests* (Chicago: Science Research Associates, Inc., 1954).

²⁶ *Wing Standardized Tests of Musical Intelligence* (London: National Foundation for Educational Research in England and Wales, 1939).

visual stimuli, to match sounds as in rhyming, and the like. A "job analysis" of beginning reading has led to hypotheses concerning types of items that might have predictive validity. Those items that have actually shown predictive validity in several different situations have been retained. A well-designed reading-readiness test does seem to have higher predictive validity than a test of general mental ability. An additional advantage is that the results of reading-readiness tests are more easily discussed with parents and do not lead to premature generalizations about the child's general intelligence.

The *Gates Reading Readiness Tests* and the *Metropolitan Readiness Tests* are probably the most widely used group tests of reading readiness. The latter includes subtests on readiness for early work in arithmetic. Several other tests, together with information on reviews, are listed in Appendix A. In interpreting the results of these readiness tests, the teacher should consider the test scores in combination with other data indicative of the child's readiness for first-grade learning experiences.

PROGNOSTIC TESTS IN FOREIGN LANGUAGE The renewed interest in the teaching of foreign language has led to the development of the first prognostic test in foreign language published in many years. In 1958, Carroll and Sapon published their *Modern Language Aptitude Test*. Based on a five-year research study conducted at Harvard University, this test has been shown to have predictive validity for the student's progress in learning almost any modern or classical language. Since the test is only moderately correlated with intelligence, it provides "new information" to the counselor and may be used in combination with intelligence test data in developing two-variable expectancy charts from local data.

PROGNOSTIC TESTS IN MATHEMATICS Two different approaches have been used in prognostic testing in mathematics. Some tests, such as the *Orleans Algebra Prognosis Test*, measure the student's speed and accuracy in learning material similar to that which he will encounter in the course. Other tests, such as the *Iowa Algebra Aptitude Test*, constitute an inventory of the student's achievement in the underlying skills (that is, arithmetic computation, manipulation of numerical series, computations involving abstract concepts, and solution of problems involving dependence and variation). Since multiscore tests of general mental ability and vocational-aptitude test batteries usually include a subtest on numerical ability, the use of special prognostic tests in mathematics may decline.

PROGNOSTIC TESTS IN SHORTHAND Prognostic tests have been used to considerable advantage in measuring aptitude for learning shorthand skills. The need for predicting success in shorthand is perhaps greater than the

need in any other secondary school subject, because (1) the mortality in shorthand courses is very high, (2) shorthand skills have limited value for students who do not use them vocationally and soon deteriorate if they are not maintained through regular practice, (3) many students with low general intelligence need to be guided away from a stenographic goal to a more suitable vocational choice, and (4) students' cumulative records contain little relevant information.

Tests of shorthand aptitude are listed in Appendix A. The Turse and the ERC tests seem to have high face validity as well as predictive validity. That is, students can more readily accept their test scores as relevant because of the obvious similarity between the abilities required in the test exercises and those required in shorthand.

PURPOSES FOR WHICH APTITUDE TESTS ARE USED

The preceding chapter sections on types of aptitude tests have revealed that a bewildering variety of aptitude tests are available. In fact, many measurement textbooks have three or more chapters on aptitude testing, with separate chapters being devoted to (1) individual tests of general mental ability, (2) group tests of general mental ability, (3) aptitude test batteries, and (4) tests of special abilities. The reason that we have grouped all these tests in a single chapter is to emphasize their common functions and to indicate the responsibility of the test user to make discriminating choices among them in terms of his purposes in testing.

Moreover, we believe that the current trend toward interpreting test data to parents and students may result in a trend away from the routine administration of tests designed to measure the construct of "general intelligence" and toward the use of a variety of aptitude tests, each being selected as most appropriate to the school's current purposes in testing. For example, a school staff might decide that:

1. A readiness test would help best in grouping and diagnosis at the first-grade level and minimize risks of premature generalizations about a child's intelligence.
2. Two tests of general mental ability should be administered during the elementary school years, with the position of the test on the spectrum of ability tests (Fig. 5.1) being affected by the characteristics of the student population and by whether or not achievement tests are routinely administered.²⁷

²⁷ If achievement tests were not routinely administered, a test emphasizing school-developed abilities such as SCAT might be desirable. On the other hand, if such achievement tests are routinely given, a test that emphasizes novel test situations and abilities that are largely learned in nonschool situations might provide more new information and help to identify children with unrecognized potential.

3. At the junior high school level a test that provides separate scores in verbal and numerical abilities might be helpful in early decisions regarding elective courses.
4. At the senior high school level an aptitude test battery with subtest scores related to future college achievement, as well as success in various occupational groups, would be more adequate and more easily discussed with students and parents than a test of general mental ability.
5. Teachers of art, music, and industrial arts might be encouraged to use special ability tests as an aid to students in self-appraisal.
6. Prognostic tests in foreign language might be used to advantage in counseling students regarding the advisability of taking regular or accelerated courses; while prognostic tests in shorthand might be administered to students beginning the commercial curriculum to help them estimate the probability of their achieving an adequate proficiency level in shorthand.

As we have already emphasized, the elementary schools are usually concerned with obtaining measures of *general* mental ability for each pupil, rather than profiles of mental abilities. If the school is concerned only with making the best assessment of the level of academic work of which a student is capable, for example, as *one* basis for grouping students, a good omnibus test of items predictive of academic achievement will do the job most efficiently. However, such a test will add little in the way of "new information" for the large number of students who are making normal progress in school work. If a test is chosen that includes subtests on nonverbal items, the predictive validity of the test for scholastic achievement will be lower than if the same amount of testing time had been devoted to items involving verbal abilities. However, new information not provided by achievement tests will be obtained, and teachers may be able to identify students who have a higher level of learning ability than was evident from a verbal test alone. In a school with a large number of children from bilingual and/or underprivileged homes, routine use of a test providing verbal and nonverbal IQ's is advisable. In other schools, it may be advantageous to use verbal tests with their greater predictive efficiency, but to provide supplementary testing with a nonverbal test for students who are retarded in their linguistic development.

High school students must make decisions concerning the advisability of attending different types of postsecondary schools, on the types of high school curricula best suited to their abilities and goals, and on the advisability of enrollment in certain specialized courses, such as shorthand. In order to make these and other decisions wisely, the student needs the most adequate data he can obtain concerning his ability to progress satisfactorily in such new experiences. The use of adequate tests may help to avoid costly trial-and-error experimentation and the psychological consequences of failure.

As the student reaches the eleventh or twelfth grade, profiles of aptitude

test data become especially valuable. Here the student faces more specific choices, of vocation or college major; and the need for reappraisal is indicated. Aptitude- and interest-test data obtained at this grade level have relatively high reliability, can be more meaningfully interpreted, and have greater significance for the more mature students.

INTERPRETATION OF RESULTS FROM APTITUDE TESTS

Interpretation of Results from Scholastic Aptitude Tests

The results of scholastic aptitude tests have usually been interpreted in terms of MA's and IQ's. A child's MA on a specific test of scholastic aptitude is the average age of children in the norming sample who did as well on the test as he did.

If a six-year-old does as well on a scholastic aptitude test as the average eight-year-old in the norming sample, his MA is 8.0. The problems involved in interpreting age scores are considered in Chapter 2. If the test is quite a difficult one that contains many items suitable for eight-year-olds, we may be able to infer that this child has a level of mental development typical of the average eight-year-old. Certainly we know that he will be able to handle more advanced types of learning experiences, and will probably progress more rapidly, than the typical first-grader. Obviously, however, he has not developed many of the concepts and skills needed to do third-grade work (the grade in which most eight-year-olds are enrolled). The MA, however, despite its limitations, is the best *single* index of readiness for different levels of difficulty in intellectual activities.

A child's IQ, which is a measure of his *rate* of mental development, was traditionally obtained by the following ratio formula:

$$IQ = \frac{MA}{CA} \times 100$$

In this case, we would obtain:

$$IQ = \frac{8}{6} \times 100 = 133$$

This six-year-old's rate of development appears to be at the rate of one and one-third years of growth in scholastic aptitude for each year of chronological age. If we compare his MA's in successive elementary school grades with those for another child with an IQ of 100, we will find that the difference between their mental ages will become increasingly large. If their IQ's on successive tests remain the same, the difference in MA by the time

they reach the seventh grade at age 12, will be four years, rather than two years, as at present. Since the child with an IQ of 133 has a much higher rate of development than the average child, the disparity between their levels of scholastic aptitude, or expected achievement in intellectual tasks, will increase.

If two children each had an MA of 6.0 when they entered the first grade, we cannot assume that each of them has a MA of 7.0 a year later when they enter the second grade. The child with a high IQ (or rate of mental development) will have grown more than a year in mental age, while the dull child will have grown less. Hence, unless a test has been very recently given, the MA must be estimated by the following version of the IQ formula:

$$MA = \frac{IQ (CA)}{100}$$

For the child we have been considering with an IQ of 133, the computation for age seven would be:

$$MA = \frac{133 (7)}{100} = 9.3$$

Ordinarily 84 months would have been substituted for the age, so that the MA would be found in months, rather than decimal fractions of a year.

DISADVANTAGES OF THE RATIO METHOD OF COMPUTING IQ'S As we have explained, the IQ has traditionally been obtained by a ratio formula. If the student refers back to the different types of number systems in Chapter 2, page 63, he will find that when we divide one number by another, we assume that the measurement units are equal and that the scale has a meaningful zero point. Reference to the section on age norms will remind the student that the mental-age unit does not represent equal values at all ages. It is gratifying, therefore, to note that as mental ability tests are revised, an increasing number of them are abandoning the ratio IQ.

The ratio method also introduced difficulties into the computation of intelligence quotients for unusually bright or unusually dull children. For example, Peter, a 10-year-old pupil in the fifth grade, does as well as the average 15-year-old pupil in the norming population on a *fifth-grade intelligence test*. His MA would, therefore, be 15 and his ratio IQ would be 150. It is highly doubtful, however, that Peter would achieve as well as the average 15-year-old pupil on a test designed for this higher level of maturity. The use of the ratio method thus involves assumptions that are not valid for students of high school age or for those pupils in the elementary schools who are extremely bright or extremely dull. The deviation

IQ, a normalized standard score, is now becoming more widely used. Since the 1960 revision of the Stanford-Binet has changed to the deviation IQ, the three leading individual tests all use this method of obtaining IQ's.

According to this procedure, the score earned by each student on an intelligence test is simply compared with the scores of other students of *his own age*. His position is ascertained in a normal distribution for his own age group, and that position (actually a standard score) is translated into an intelligence quotient. According to this plan, Peter's score would be compared not with that of the average 15-year-old but with those of all 10-year-olds in the norming population. If he excelled 98 percent of his own age group, his IQ would be 130. This method of computing deviation IQ's is used in the Pintner and Otis series of intelligence tests and in the *Terman-McNemar Group Test of Mental Ability*. The relationship of deviation IQ's to standard scores and percentile ranks is clarified in Appendix C.

CONSTANCY OF THE IQ When the IQ was defined as rate of mental growth, the assumption was implicit that a measurable rate of growth did exist for each child, and that it was reasonably constant. Such an assumption, however, did not carry the implication that IQ's obtained from different tests and in different circumstances would agree perfectly. Not only do the samplings of intellectual abilities differ from test to test, but many other factors contribute to variability in test results.

It is well to recognize that an IQ is derived from the score obtained on a *single* intelligence test. It is not a *direct* measure of an individual's rate of mental growth but a converted score, based on comparing one measure of the student's performance on intellectual tasks with the performance of students in the norming sample on these same tasks. One cannot infer that, if the IQ obtained on a later intelligence test is higher, the individual has become brighter since the time of his previous testing.

The concepts of error variance and the standard error of measurement must be taken into account in any interpretation of intelligence test results. Investigations show that the average change in IQ on repeated *individual* intelligence tests is approximately 5 points, that 20 percent of IQ's change 10 or more points, and that approximately 1 percent change 20 or more points. However, if the tests are group tests, and especially if the test used at the later testing is not the same as that used earlier, the differences may be even greater.²⁸

²⁸ Florence L. Goodenough, "New Evidences on Environmental Influence on Intelligence," *Intelligence: Its Nature and Nurture*, 39th Yearbook, Part I (Chicago: National Society for the Study of Education, 1940), p. 358.

THE ASSUMPTION OF EQUAL OPPORTUNITY TO LEARN A student's intelligence or scholastic aptitude is estimated by obtaining a sampling of his behavior in a test situation in which the items have been especially selected to reveal his learning ability. In selecting items for intelligence tests, authors try to include items that are *equally unfamiliar* to students (described as *novel situations*) or *equally familiar* to students (in the sense that all students have had "equal opportunity to learn" the material included).

Allison Davis has challenged the assumption that typical intelligence tests include materials that students of different socioeconomic classes have had equal opportunity to learn. He contends that students of the middle and upper classes have had greater opportunity and certainly greater motivation to extend their vocabulary, to learn abstract symbols, and to study reasoning problems in arithmetic than have students of the lower socioeconomic classes.²⁹ It is certainly evident that such a word as "sonata" in the vocabulary section of an intelligence test would be known by relatively more students from the higher socioeconomic classes than by those from the lower classes. Certainly an attempt should be made to include test items that do not give special advantage to children reared in the upper or middle classes, or to children reared in urban vs. rural environments.

If one is concerned only with the intelligence test as a measure of the child's readiness for learning activities, and as a predictor of success in schoolwork, he can ignore such criticisms as Davis has made. That is, he can say that the child whose culture has limited his vocabulary development will obtain a somewhat lower scholastic aptitude score but will also tend to score lower on the criterion of success in school. Certainly if we omit from scholastic aptitude tests all items that discriminate against the culturally disadvantaged child, our test will be a less adequate predictor of school achievement. Research on the SAT test in Chapter 5 led to the conclusion that the unfairness lay not in the test, but in the culture. In other words, we should be cautious about the inferences we make from intelligence tests, limiting our inferences to those concerning the child's probable success in school work, rather than overgeneralizing about his general intelligence or his innate ability.

Davis and Eells attempted to develop a series of intelligence tests for the elementary grades that would minimize such cultural bias. Anastasi concluded, on the basis of a summary of 30 research studies, that Davis and Eells, in their attempt to develop a culture-fair test, had sacrificed predictive validity without eliminating cultural bias. That is, predictive

²⁹ Allison Davis and Kenneth Eells, *Davis-Eells Test of General Intelligence or Problem-Solving Ability* (New York: Harcourt, Brace & World, Inc., 1953). For research data on this problem, see Allison Davis and Others, *Intelligence and Cultural Differences* (Chicago: University of Chicago Press, 1951).

validity coefficients with both achievement tests and teachers' ratings were uniformly lower than for conventional intelligence tests. This loss in predictive validity might have been justifiable if the authors had achieved their goal of developing a test that would more adequately measure the learning ability of lower-class children. However, several studies have revealed that lower-class children perform as poorly on the Davis-Eells tests as on conventional intelligence tests.³⁰

In studying the intelligence of people from quite different cultures (as in cross-cultural research studies), tests that rank on the lower end of the continuum on verbal-educational loading should be used (see Fig. 5.1). Such tests, however, will tend to have low predictive validity for most criteria of success in educational or vocational activities. For the student who wishes to study the problem of cultural bias in intelligence testing, a number of references are listed in the first section of the Selected References for this chapter.

VARIATIONS IN TEST CONTENT Scholastic aptitudes tests can sample evidences of intelligent behavior only as they are revealed in student responses to the items selected for each specific test. Content varies considerably from test to test. Some tests place a greater premium on reading ability and vocabulary; others involve more items on numerical skills and understandings; while still other tests include a large percentage of items requiring perceptual speed and memory. A student's relative rank in his age group, and hence his IQ, will vary as a test includes more or less of the types of intellectual tasks on which he does best. Moreover, the types of test items that can be included in an intelligence test for kindergarten and first-grade children obviously differ from those that can be included in a test for older students, in which reasoning, judgment, and other higher mental processes can be more adequately tested.

A student's IQ should always be interpreted as his converted score on a specific intelligence test given at a specific time. That is, instead of saying that Jerry's IQ is 120, one should say that Jerry has an IQ of 120 on a specific test (for example, the *Pintner General Ability Test, Verbal Series, Intermediate*, taken in the eighth grade). If Jerry obtains a lower IQ on a test of performance-type items or a higher IQ on a test placing a great premium on speed, the results are not necessarily inconsistent or contradictory.

OTHER REASONS FOR VARIABILITY IN INDIVIDUAL IQ'S Although differences in the *content* of intelligence tests constitute one of the major causes, there are several other reasons for the variability in results:

³⁰ Anne Anastasi, *Psychological Testing* (New York: The Macmillan Company, 1961, p. 267).

1. The use of tests that have low reliability.
2. Changes in the individual from one testing occasion to another, for example, changes in the subject's physical condition or emotional adjustment, including *his attitude toward the test situation*.
3. Unreliability of test results for young children because of variations in their motivation, attentiveness when directions are given, and the like.
4. The ratio method of computing the intelligence quotient (as explained above).
5. Differences between the norming samples on which tests are standardized.

Although many test publishers are now taking great care in selecting their norming samples, norming problems are still a factor in the variability of intelligence quotients. An illustrative study on this last factor in test variability indicated that the *Otis Quick-Scoring Mental Ability Tests* were standardized on a norming population that was unusually homogeneous, with an *SD* of only 10 IQ points as compared with 13 for the Pintner. Such differences in *SD* result in these tests giving almost identical IQ's when the results are near 100 but increasingly divergent results as the IQ's differ from the mean. For example, an Otis IQ of 100 is equal to a Pintner IQ of 99; however an Otis IQ of 120 is equal to a Pintner IQ of 123; and an Otis IQ of 135 is equal to a Pintner IQ of 140.³¹

Interpretation of Results from Aptitude Test Batteries

The interpretation of aptitude test scores in counseling interviews is considered in Chapter 17. Here we consider such significant problems as the probable effects of the interpretation of test results on the student's concept of self and his aspirations for the future. The following section is concerned chiefly with those problems of interpretation that grow out of the limitations of the tests themselves.

THE NEED FOR CONSIDERING OTHER EVIDENCE Perhaps the most important single principle in aptitude test interpretation is that the test results must be considered *in the context of all other information* on the

³¹ Roger T. Lennon, "A Comparison of Results of Three Intelligence Tests," *Test Service Notebook*, No. 4 (New York: Harcourt, Brace & World, Inc., n.d.). Copies available on request. The three tests studied were *The Otis Quick-Scoring Mental Ability Tests: Gamma Test*; *The Pintner General Ability Tests: Verbal Series, Advanced Test*; and *The Terman-McNemar Test of Mental Ability*. A publication of The Bureau of Pupil Guidance, Chicago Public Schools, entitled "Equivalence of Intelligence Quotients of Five Group Intelligence Tests," gives IQ equivalents for five intelligence tests: *The Lorge-Thorndike*, *Otis Quick-Scoring*, *California Mental Maturity*, *Kuhlmann-Anderson*, and *Pintner*. A summary table of the results is given in Irving Lorge and Robert Thorndike, *Technical Manual, The Lorge-Thorndike Intelligence Tests* (Boston: Houghton Mifflin Company, 1962).

student's abilities (available in school records or supplied by students and parents). As has been mentioned, *achievement* test results are good indicators of aptitude, provided that the student has had typical opportunities for learning and motivation to learn. The student's *pattern of marks* in various subject fields may provide useful clues as to both aptitudes and interests.

Achievement in exploratory courses (general shop, general music, and the like), work experience, extracurricular activities, and hobbies all reflect the student's abilities and interests. However, caution must be used in making predictions regarding probable vocational success on the basis of students' interest and achievement in school activities. Participation in A Cappella Choir or Senior Band, for example, may not be predictive of success in music vocations; a role in a school play may not indicate talent sufficient to succeed in the highly competitive field of dramatics; nor does the ability to sketch costumes necessarily qualify a student for a career in fashion illustration.

Hobbies and extracurricular activities are especially valuable as indicators of *interest* in helping the student to choose among several occupations, all of which have about the same requirements with respect to general or special mental abilities. They are less valuable, however, as indicators of *aptitude*, and may even be misleading for students desiring to enter the professions, which require rigorous preservice training, or vocations in the fine arts, which are highly competitive. Nevertheless, for all students, self-evaluation of try-out experiences in courses and activities help to provide a more adequate basis for interpreting the more objective test data.

REPORTING APTITUDE-TEST RESULTS TO STUDENTS Although this subject is considered at greater length in Chapter 17, it is appropriate at this point to emphasize that (1) a student's profile of aptitude test results cannot be drawn unless the tests to be compared are normed on the same population or on comparable norming groups; (2) the profile should be drawn on a scale that reflects the nature of percentile norms, as in Figure 6.2; (3) the aptitude test data should be summarized together with achievement test results, a record of marks by subject field, interest-inventory findings, and other relevant data; and (4) the presentation of findings should be handled in such a way as to promote maximum student interest and growth in self-appraisal and self-direction.

THE FORECASTING OF SUCCESS AND FAILURE If a girl achieves high percentile ranks on the clerical speed and accuracy and the language usage subtests of the *Differential Aptitude Tests*, the counselor can justifiably inform her that she has a high probability of success in courses preparing for stenography. However, there are so many unmeasured factors that

affect vocational success that the *degree* of her success in a stenographic position cannot be predicted. If, on the other hand, the student has made *low* scores on the clerical-aptitude tests, and has had little success in related school experiences, the counselor can indicate with some assurance that she has little likelihood of success in the prerequisite training program or on the job. Similarly, low scores on tests of music or art aptitude or manual dexterity, combined with supporting data from school or job experiences, would justify counseling students against entering vocations that require such abilities. High scores, however, would by no means be sufficient to justify a prediction of a high degree of vocational success in those occupations.

Aptitude tests can be valuable in indicating the fields in which a student is *unlikely* to succeed. There is great need for more research, however, on the minimal or critical score, which would indicate minimum aptitude in critical abilities for various occupations. Research studies by the United States Employment Service have indicated that unique patterns of abilities for specific vocations do not, in fact, exist. Rather, for each job there appears to be only a set of minimal requirements. In other words, if an individual has at least the minimal level of ability in the various component characteristics required in a vocation, the probability of his success may depend chiefly on personality and motivational factors. In the use of the *General Aptitude Test Battery* (described in Table 6.4), critical, or "cut-off," scores³² are provided, which constitute a major basis for interpretation.

THE NEED FOR MORE EXTENSIVE VALIDATION DATA The publishers of the *Differential Aptitude Tests*, the *Flanagan Aptitude Classification Tests*, and other new aptitude test batteries are committed to research programs involving (1) the norming of these tests for occupational groups and (2) study of the relationship of aptitude test scores to educational and vocational success. The latest supplement to the DAT manual summarizes a large number of such research studies on this battery.

In order to obtain maximal value from aptitude tests, the counselor should have at hand, for each aptitude test, percentile or standard-score norms for many relevant occupational groups and for students in training for relevant occupations. The counselor should be able to tell the student how his scores compare with those of college or occupational groups with which he will have to compete.

Data on the relationship of aptitude-test scores to students' marks in various high school and college curricula are also significant. In fact, local expectancy tables for predicting grades in college-preparatory and voca-

³² The 33d percentile point, arbitrarily taken as distinguishing the "less able" from the "more able" workers.

tional subjects are very useful in interpreting aptitude test scores to students.³³

APTITUDE TEST RESULTS AND DIFFERENTIAL PREDICTION The chief concern of the student and his parents is with *differential* prediction—"Will I do *better* in law than in architecture?" "Am I *better* fitted for teaching than for nursing?" Aptitude test scores do not provide direct answers to such questions. The counselor, however, can assist the student in summarizing and interpreting the available data. Techniques for differential prediction are presented in Chapter 17.

Bennett and Doppelt point out that if the *differences* between student results on two tests are to be significant, the intercorrelation of the tests should not exceed .60, with lower values desirable. They have developed a nomograph by which one can use the reliability coefficients and intercorrelations of pairs of tests to determine the probability of finding reliable differences between pairs of student scores.³⁴

The counselor needs aptitude tests that have the greatest promise for differential prediction because they (1) are the most reliable tests that can be obtained; (2) are accompanied by a wealth of validation data that help the student to compare himself with various groups; (3) show, by a pattern of low intercorrelations, that they measure abilities that are not closely interrelated; and (4) provide data that help guidance workers to interpret the statistical significance of *differences* between scores.

SUMMARY STATEMENT

Aptitude tests are designed to measure the individual's ability to progress in learning activities of some specified type. The validity of an aptitude test is judged in terms of the value of its scores in predicting future performance.

Some aptitude tests are specially designed to measure aptitude for some subject or vocation, that is, to measure a combination of abilities related to future success in that subject or vocation. Other tests are designed to measure the individual's performance level in one or more discrete, unitary abilities, such as verbal comprehension or perceptual speed, which affect an individual's performance in many different subjects or occupations. In preparing aptitude tests of either type, authors attempt to include tasks which are (1) equally

³³ For suggestions on the summarization of local data, see "Expectancy Tables—a Way of Interpreting Test Validity," *Test Service Bulletin No. 38* (New York: The Psychological Corporation, 1949). Illustrative expectancy tables are given in Chapters 4 and 17.

³⁴ George K. Bennett and Jerome E. Doppelt, "Evaluation of Pairs of Tests for Guidance Use," *Educational and Psychological Measurement*, vol. 8 (Autumn 1948), pp. 319-323.

unfamiliar to all examinees or (2) equally familiar (in the sense that all examinees have had approximately equal opportunity to learn them).

Although no attempt was made to describe all the leading aptitude tests, illustrations of all the major types were given. Comparisons were made in the text and in summary tables that were designed to help students understand similarities and differences between: (a) an aptitude test battery and a multiscore test of mental abilities, (b) two leading individual intelligence tests, (c) two leading vocational aptitude tests, and (d) a performance and paper-and-pencil test of the same special aptitude (spatial visualization).

Instead of routinely administering tests of general intelligence at several grade levels, school authorities might well choose from a variety of aptitude tests those that provide the most valuable information in making the types of decisions needed at each level.

The results of scholastic aptitude tests have usually been interpreted in terms of MA's and IQ's. The individual's mental age can be interpreted in terms of his *level* of mental maturity or his readiness to undertake learning tasks of a certain level of complexity. The IQ indicates the student's *rate* of mental development and is useful in predicting his rate of progress in learning activities appropriate to his mental maturity level. The limitations of age scores and of the ratio method of computing IQ's were considered.

Teachers are rightfully concerned about the range in the IQ's obtained for an individual during his school years. Variability in intelligence-test results can be reduced, in part, by a systematic program of administering intelligence tests at least biennially, careful selection of tests, desirable conditions of administration, and retesting of certain students. Teachers should recognize the need for supplementing verbal tests by nonverbal tests when there is a reading handicap, and of using individual tests when there is reasonable doubt as to the ability or willingness of a student to do his best work on a group test.

In studying intra-individual differences in aptitudes, it is essential that tests of the different aptitudes be normed on the same or comparable norming samples. Although the use of such tests in counseling is considered in Chapter 17, several principles of interpretation were emphasized in this chapter: (1) Other evidence, such as achievement test data, students' marks, and the like, should be considered. (2) Caution should be exercised in reporting aptitude test results to students. (3) Aptitude tests forecast failure more reliably than success. (4) More extensive validation data are needed for use in test interpretation. (5) Aptitude tests results are frequently inadequate for differential prediction.

SELECTED REFERENCES

Cultural Factors Affecting Aptitude Test Scores

- ANASTASI, ANNE, "Some Implications of Cultural Factors for Test Construction," *Proceedings of the 1949 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 1950, pp. 13-17.
- COLEMAN, WILLIAM, AND ANNIE W. WARD, "A Comparison of Davis-Eells and Kuhlmann-Finch Scores of Children from High and Low Socio-Economic Status," *Journal of Educational Psychology*, vol. 46 (December 1955), pp. 463-469.

- DARCY, NATALIE T., "A Review of the Literature on the Effects of Bilingualism upon the Measurement of Intelligence," *Pedagogical Seminary and Journal of Genetic Psychology*, vol. 82 (March 1953), pp. 21-57.
- EELLS, KENNETH, AND OTHERS, *Intelligence and Cultural Differences*. Chicago: University of Chicago Press, 1951.
- JONES, W. R., "A Critical Study of Bilingualism and Non-Verbal Intelligence," *British Journal of Educational Psychology*, vol. 30 (February 1960), pp. 71-77.
- LEVINSON, BORIS M., "Subcultural Variations in Verbal and Performance Ability at the Elementary School Level," *Journal of Genetic Psychology*, vol. 97 (September 1960), pp. 149-160.
- NOLL, VICTOR H., "Relation of Scores on Davis-Eells Games to Socio-Economic Status, Intelligence Test Results, and School Achievement," *Educational and Psychological Measurement*, vol. 20 (Spring 1960), pp. 119-129.

Other Factors Affecting Validity of Aptitude Test Scores

- FRANKEL, EDWARD, "Effects of Growth, Practice, and Coaching on Scholastic Aptitude Test Scores," *Personnel and Guidance Journal*, vol. 38 (May 1960), pp. 713-719.
- FRENCH, JOHN W., AND ROBERT E. DEAR, "Effect of Coaching on an Aptitude Test," *Educational and Psychological Measurement*, vol. 19 (Autumn 1959), pp. 319-330.
- HOLLOWAY, H. D., "Effects of Training on the SRA Primary Mental Abilities (Primary) and WISC," *Child Development*, vol. 25 (December 1954), pp. 253-263.
- MAURER, KATHERINE M., *Intellectual Status at Maturity as a Criterion for Selecting Items in Preschool Tests*. Minneapolis, Minn.: University of Minnesota Press, 1946.
- SUPER, DONALD E., AND JOHN O. CRITES, *Appraising Vocational Fitness by Means of Psychological Tests*. New York: Harper & Row, Publishers, Inc., 1962, Chapters 5, 6, 10-14.

Stability of Aptitude Test Scores

- BAYLEY, NANCY, "Data on the Growth of Intelligence between 16 and 21 Years as Measures by the Wechsler-Bellevue Scale," *Journal of Genetic Psychology*, vol. 90 (March 1957), pp. 3-15.
- BRADWAY, KATHERINE P., CLARE W. THOMPSON, AND R. B. CRAVENS, "Preschool IQ's After Twenty-five Years," *Journal of Educational Psychology*, vol. 49 (October 1958), pp. 278-281.
- MEYER, WILLIAM J., "The Stability of Patterns of Primary Mental Abilities among Junior High and Senior High School Students," *Educational and Psychological Measurement*, vol. 20 (Winter 1960), pp. 795-800.
- TYLER, LEONA E., "The Stability of Patterns of Primary Mental Abilities among Grade School Children," *Educational and Psychological Measurement*, vol. 18 (Winter 1958), pp. 769-774.

Uses of Aptitude Tests

- DAVIS, FREDERICK B., *Utilizing Human Talent* (Washington, D.C.: American Council on Education, 1947).

- GHISELLI, E. E., *Measurement of Occupational Aptitude* (Berkeley, Calif.: University of California Press, 1955).
- Identification and Guidance of Able Students* (Washington, D.C.: American Association for the Advancement of Science, 1958).
- STROUD, J. B., "The Intelligence Test in School Use: Some Persistent Issues," *Journal of Educational Psychology*, vol. 48 (February 1957), pp. 77-85.
- SUPER, DONALD E., ed., *The Use of Multifactor Tests in Guidance* (Washington, D.C.: American Personnel Guidance Association, 1958). Reprinted from *Personnel and Guidance Journal*, vol. 35 (September 1956), pp. 9-51.

Prognostic Testing in Subject Areas

- BANHAM, KATHARINE M., "Maturity Level for Reading Readiness—A Check List for the Use of Teachers and Parents as a Supplement to Reading Readiness Tests," *Educational and Psychological Measurement*, vol. 18 (Summer 1958), pp. 371-375.
- CARROLL, JOHN B., "A Factor Analysis of Two Foreign Language Aptitude Batteries," *Journal of General Psychology*, vol. 59 (July 1958), pp. 3-19.
- FARNSWORTH, PAUL R., *Musical Taste: Its Measurement and Cultural Nature*, Stanford University Publications in Education-Psychology, vol. II, No. 1. (Stanford, Calif.: Stanford University Press, 1950), Chapters 3-4.
- HORN, CHARLES A., AND LEO F. SMITH, "The Horn Art Aptitude Inventory," *Journal of Applied Psychology*, vol. 29 (October 1945), pp. 350-355.
- SALOMAN, ELLEN, "A Generation of Prognosis Testing," *Modern Language Journal*, vol. 38 (October 1954), pp. 299-303.
- WING, H., "Tests of Musical Ability and Appreciation: An Investigation into the Measurement, Distribution and Development of Musical Capacity," *British Journal of Psychology, Monograph Supplement* 27, 1948. (Chicago: University of Chicago Press, 1949).

General and Theoretical References

- BURT, CYRIL, "The Differentiation of Intellectual Ability," *British Journal of Educational Psychology*, vol. 24 (June 1954), pp. 76-90.
- DYER, HENRY S., "A Psychometrician Views Human Ability," *Teachers College Record*, vol. 61 (April 1960), pp. 394-403.
- GUILFORD, J. P., "Three Faces of Intellect," *American Psychologist*, vol. 14 (August 1959), pp. 469-479.
- HUMPHREYS, LLOYD G., AND PAUL L. BOYNTON, "Intelligence and Intelligence Tests," *Encyclopedia of Educational Research*. New York: The Macmillan Company, 1950, pp. 600-612.
- KETTNER, NORMAN W., J. P. GUILFORD, AND PAUL R. CHRISTENSEN, "A Factor-Analytic Study across the Domains of Reasoning, Creativity and Evaluation," *Psychological Monographs*, vol. 73, No. 9, Whole No. 479 (Washington, D.C.: American Psychological Association, 1959).
- LEVINE, ABRAHAM S., "Aptitude Versus Achievement Tests as Predictors of Achievement," *Educational and Psychological Measurement*, vol. 18 (Autumn 1958), pp. 517-525.
- VERNON, P. E., *The Structure of Human Abilities* (New York: John Wiley and Sons, Inc., 1950):

References re Specific Aptitude Tests

- BENNETT, G. K., AND OTHERS, *Counseling from Profiles: A Casebook for the Differential Aptitude Tests* (New York: The Psychological Corporation, 1953).
- BURKE, HENRY R., "Raven's Progressive Matrices: A Review and Critical Evaluation," *Journal of Genetic Psychology*, vol. 93 (December 1958), pp. 199-228.
- GWYNNE-JONES, H., "The Evaluation of the Significance of Differences between Scaled Scores on the WAIS: Perpetuation of a Fallacy," *Journal of Consulting Psychology*, vol. 20 (August 1956), pp. 319-320.
- HARRIS, DALE B., *Measuring the Psychological Maturity of Children: A Revision and Extension of the Goodenough Draw-A-Man Test*. (New York: Harcourt, Brace & World, Inc., 1961).
- JONES, LYLE V., "Primary Abilities in the Stanford-Binet, Age 13," *Journal of Genetic Psychology*, vol. 84 (March 1954), pp. 125-147.
- LITTELL, WILLIAM M., "The Wechsler Intelligence Scale for Children: Review of a Decade of Research," *Psychological Bulletin*, vol. 57 (March 1960), pp. 132-156.
- MCNEMAR, QUINN, "On WAIS Difference Scores," *Journal of Consulting Psychology*, vol. 21 (June 1957), pp. 239-240.
- MEYER, WILLIAM H., AND A. W. BENDIG, "A Longitudinal Study of the Primary Mental Abilities Test," *Journal of Educational Psychology*, vol. 52 (February 1961), pp. 50-60.
- SCHUTZ, RICHARD E., "Factorial Validity of the Holzinger-Crowder Uni-Factor Tests," *Educational and Psychological Measurement*, vol. 18 (Winter 1958), pp. 873-875.
- SHARP, H. C., AND L. M. PICKETT, "The General Aptitude Test Battery as a Predictor of College Success," *Educational and Psychological Measurement*, vol. 19 (Winter 1959), pp. 617-623.
- TERMAN, LEWIS M., AND MAUD A. MERRILL, *Measuring Intelligence*. Boston: Houghton Mifflin Company, 1959.
- United States Department of Labor, Bureau of Employment Security, *Guide to the Use of General Aptitude Test Battery: Section III. Development*. (Washington, D.C.: Government Printing Office, 1958).
- WECHSLER, D., *The Measurement and Appraisal of Adult Intelligence*, 4th ed. (Baltimore: The Williams and Wilkins Company, 1958).

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. Summarize early developments in the mental testing of individuals. What contributions were made by Binet? By Terman? When were group intelligence tests developed and what factors led to their development?
2. How does the basic approach to the construction of intelligence tests differ from the basic approach used in constructing achievement tests?
3. Study the reviews of one aptitude test from one of the *Mental Measurements Yearbooks*. Summarize the evidence given concerning the concurrent and predictive validity of that test.
4. Examine the manuals for two leading aptitude test batteries. Summarize the research data presented on the construction and validation of each test.

5. Compare subtests from two aptitude test batteries for a single aptitude, such as spatial relations or verbal comprehension. Do the tests involve essentially the same task for the student? How do they compare with respect to working time, clarity of instructions, norms available, validation data, and the like?

6. For what purposes are tests of single aptitudes used? What are the advantages and disadvantages of using two or more of these tests, in comparison with an aptitude test battery measuring the same abilities?

7. Assume that you are a high school counselor. Prepare notes for a talk to teachers on the interpretation of results on aptitude test batteries.

8. In many cases a child's language IQ is 10 or more points below his non-language IQ. What are the possible reasons for such large differences?

9. What are the uses for individual intelligence tests in a school in which group intelligence tests are regularly administered?

10. Why might a specific test of scholastic aptitude have low validity for some children?

11. Distinguish between tests of manual dexterity and mechanical aptitude. Cite an example of a test of each type. How is each type used in vocational guidance?

12. What is experiential maturity and why is it an important factor in reading readiness in the first grade? What informal methods can a teacher use to obtain evidence on a pupil's readiness for reading?

13. In the following list of educational problems, indicate whether the IQ or the MA should be given greater consideration:

- a. Estimating readiness for reading in the first grade
- b. Determining the optimum grade placement for a student entering elementary school
- c. Estimating the amount of practice needed to attain mastery of certain arithmetic skills
- d. Selecting the candidates for a special class for the gifted
- e. Classifying students within a single grade level into ability groups.

14. List the criteria frequently used to validate an intelligence test, and evaluate the validity and the reliability of each criterion.

15. What significant factors in musical aptitude are not measured by the Seashore tests? What are the values and limitations of this battery as a predictor of accomplishment in music?

16. Describe and evaluate one prognostic test in your major subject field. Why is it important to have local validation data for such tests?

17. Prepare a summary statement concerning the advantages and disadvantages of prognostic testing in foreign languages, mathematics, or some other subject field.

The Measurement of Interests and Attitudes

In the remaining chapters of Part II, we turn our attention to tests of interests, attitudes, and personality traits, in which we attempt to measure the *typical* performance of individuals, rather than their *maximum* performance. With tests of this type, test results may vary with the examinee's perception of the test situation. In some situations, the examinee may frankly report his typical attitudes or behaviors as he perceives them. On other occasions, he may modify his answers in an attempt to make a good impression.

THE NATURE OF INTERESTS

During the past decade the use of interest inventories in secondary schools has markedly increased. Many interest inventories can be self-administered and self-scored; interest-inventory results intrigue students (without threatening them, as aptitude-test results may do); and the interpretation of interest profiles *appears* to be a process that can be safely attempted in group guidance classes by guidance teachers and counselors who have had training for such work. The fact that an interest inventory merely categorizes his own responses helps the student to interpret his scores as a mirror of his own reactions, rather than a mysterious dictum from some authority.

A person's interests are the product of interaction between (1) inherited bases of ability and temperament and (2) many environmental factors, notably the opportunities he has had for pursuing certain interests and the value placed on their development by persons whose approval he values. In some instances, a young child's aptitudes result in his receiving satisfac-

tion and approval for his successes; in other cases, early failures result in discouragement or early successes are met with indifference or disapproval.

Approaches to the Identification of Interests

Super and Crites¹ distinguish four major interpretations of the term "interest," associated with four methods of obtaining data on student interests.

1. *Expressed interest*—the verbal profession of interest in an activity or occupation. The student simply expresses a liking, or indicates his dislike, for a particular activity or vocation. The significance of such expressions of interest varies with the maturity and experience of the individual. In some cases, expressed interests represent temporary whims or fantasies.
2. *Manifest interest*—as evidenced by participation in an activity or occupation. A boy who is a radio "ham," a girl who is active in the dramatics club, a student who does sports reporting for the school paper are *manifesting* their interests through actual *participation*. Manifest interests tend to be more stable than expressed interests since they are based on actual experience.

This approach to the identification of interests, however, has serious limitations. Sometimes the student participating in an activity is more interested in its concomitants or by-products than in the activity itself. For example, the girl in the dramatics club may be more interested in the social contacts and prestige to be derived from club membership than in dramatics as an art or form of self-expression. She may later seek these same goals through other activities.

The manifestation of interests may be limited by financial considerations or other environmental factors. Hence, although manifest interests provide clues to possible educational and vocational goals, the absence of a specific interest may reflect only lack of environmental opportunity to develop that interest.

3. *Tested interest*—as measured by *objective tests* of vocabulary or other information rather than an *inventory* of reported interests. The use of such tests as the *Michigan Vocabulary Profile Test* as a measure of interest is based on the assumption that a stable interest results in an accumulation of relevant information and a corresponding growth in specialized vocabulary.
4. *Inventoried interest*—as measured by lists of activities or occupations to which the student responds by an expression of liking or preference. Such inventories superficially resemble questionnaires on *expressed interests*. However, many activities that have an indirect relationship to vocational choice are included. In answering the inventory items, the examinee records a series of self-perceptions that are summarized in such a way as to reveal their similarity to those of workers in different occupations (or to students of his sex and grade level). The scores of each student can be interpreted (by means of norms) as reflecting a *pattern* of relatively high or low interests in various fields.

¹ Donald E. Super and John O. Crites, *Appraising Vocational Fitness by Means of Psychological Tests* (New York: Harper & Row, Publishers, Inc., 1962), pp. 377-379.

Almost all of the interest measures that have been published and standardized for school use are of the fourth type. Experience and research suggest that interest inventories can be valuable aids in vocational guidance. The expressed interests of adolescents may be based on glamorized stereotypes of occupations, rather than an awareness of the specific activities involved. Evidence from the first three sources, however, is useful in studying the validity of published inventories and in supplementing inventory results in the counseling of individual students.

When a student's results on an interest inventory differ from his expressed interests, the counselor should not assume that the interest inventory is more valid. Instead, the counselor should investigate whether the expressed interest is a longstanding or temporary one, and whether it is based on actual experience and mature consideration. The interest inventory does have the advantage of obtaining the students' reactions to a large sampling of items and of providing, through the use of converted scores, a means of comparing the students' interests with others of his sex and age. Berdie stresses the importance of considering both expressed and inventoried interests: "As long as measured [inventoried] interests have a relevancy for vocational satisfaction and as long as self-estimated [expressed] interests play an important role in the deliberations of individuals, both types of interests must be considered."²

The Relationship of Interests and Aptitudes

It is generally accepted that interests and abilities are related—that is, a person tends to develop interests in activities that he performs easily and well and to shun those that are more difficult for him. The relationship, however, is by no means simple and clear-cut. Adkins and Kuder found only one correlation above .30 when they correlated scores on the *Chicago Primary Mental Abilities Test* and the *Kuder Preference Record*.³ In Table 7.1, which summarizes data on the interrelationships of scores on the *Differential Aptitude Test* and the *Kuder Preference Record*, very few appropriate pairings of aptitude and interest scores showed a significant relationship. In fact, when the results for the three grade levels (presented in the manual) are studied, the only appropriate relationships that were consistently significant were between boys' scores in DAT mechanical

² R. F. Berdie, "Scores on the Strong Vocational Interest Blank and the Kuder Preference Record in Relation to Self-Ratings," *Journal of Applied Psychology*, vol. 34 (February 1950), pp. 42-49.

³ D. C. Adkins and G. Frederic Kuder, "The Relation between Primary Mental Abilities and Activity Preferences," *Psychometrika*, vol. 5 (December 1940), pp. 251-262.

reasoning and their Kuder interest scores in the mechanical and scientific areas. As Wesman warns,

Experienced counselors may not need the reminder these data contain; to less experienced counselors, the results may well serve as a warning not to base counseling on interest scores without positive information with respect to the appropriate aptitudes and abilities of the student.⁴

Table 7.1

Coefficients of Correlation between Subtest Scores of the Differential Aptitude Tests and Scales of the Kuder Preference Record

Kuder Scales	Coefficients of correlation ^a with DAT tests							
	VR	NA	AR	SR	MR	CSA	SPELL.	SENT.
Mechanical	.03	.12	.14	.13	.40*	.09	-.04	-.06
Computational	.16	.22	.28*	.19	.20	.06	-.05	.00
Scientific	.46*	.27*	.36*	.13	.32*	.19	.28*	.27*
Persuasive	-.18	-.12	-.22	-.12	-.05	-.12	-.14	-.17
Artistic	-.19	-.25*	-.17	.11	-.03	-.16	.35*	-.12
Literary	.09	.02	.03	-.19	-.21	.04	.28*	.12
Musical	.13	.05	-.06	.22	-.01	.15	.04	.33*
Social Service	-.18	-.06	-.10	-.11	-.15	-.13	.06	-.16
Clerical	-.20	-.04	-.13	-.10	-.18	-.18	-.17	-.20

Source: Data presented for a group of 65 10th-grade boys in an Ames, Iowa, high school. Corresponding data for grades 11 and 12, as well as corresponding data for girls, are given in George K. Bennett, Harold G. Seashore, and Alexander G. Wesman, *Manual for the Differential Aptitude Tests*, 3d ed. (New York: The Psychological Corp., 1959), p. 75.

^a Only those coefficients followed by an asterisk are sufficiently high that the chances are 95 out of 100 that the "true r " between these tests is greater than 0.0. For the full names of DAT tests, the reader is referred to Figure 6.2.

Perhaps the most helpful statement for the counselor is that *interest* affects "the *direction of effort; ability, the level of achievement.*"⁵

We cannot infer from the low r 's in Table 7.1 and those from similar studies, that there is little relationship between interests and aptitudes. The relationship may be an exceedingly complex one. Our common sense tells

⁴ Alexander G. Wesman, "The Differential Aptitude Tests," *Personnel and Guidance Journal*, vol. 31 (December 1952), p. 169.

⁵ Super and Crites, *op. cit.*, p. 448.

us that interests are probably related to *intraindividual* differences in aptitude; and a research study by Segel⁶ found considerable support for this hypothesis.

As measurement techniques and research procedures improve, higher r 's may be found. For example, Crites⁷ discovered that individuals with average intelligence tended to score higher in "interest in technical occupations" than individuals of *either* above-average or below-average intelligence. The relationship between interest and aptitude scores in this study was a *curvilinear* one; that is, the relationship between students' scores on aptitudes and their scores in related interest areas would be represented by a curve rather than a line. Thus the degree of relationship, or the predictability of one from the other, would be underestimated by Pearson r (which assumes a linear relationship between the variables studied). The possibility of curvilinear relationships between aptitudes and certain interest areas needs to be studied further; until such researches have been completed, however, one cannot generalize concerning the relationship between interests and aptitudes in specific fields. One can be quite sure, however, that *inferences regarding level of aptitude cannot be made* on the basis of interest inventory scores.

The Relationship of Interests and Personality Traits

The processes of interest development and personality formation are so complex that it is impossible to measure the relative contributions of personality traits to the development of interests. There is considerable evidence, however, that personality factors do play a significant role in the development of vocational interests and the making of vocational choices. Neurotics are more likely than normals to be interested in "talent" and social service occupations for which they lack the requisite aptitudes.⁸ Darley found that persons in the business contact and social service fields tended to be better adjusted socially than those in literary and technical occupations.⁹

An individual's occupational choice represents a social role through which he seeks self-realization. According to Darley and Hagenah, apti-

⁶ David Segel, "Differential Prediction of Scholastic Success," *School and Society*, vol. 39 (January 20, 1934), pp. 91-96.

⁷ J. O. Crites, "Intelligence and Adjustment as Determinants of Vocational Interest Patterning in Late Adolescence." Unpublished doctoral dissertation, Columbia University, 1957, cited in Super and Crites, *op. cit.*, p. 172.

⁸ C. H. Patterson, "Interest Tests and the Emotionally Disturbed Client," *Educational and Psychological Measurement*, vol. 17 (Summer 1957), pp. 264-280.

⁹ J. G. Darley, *Clinical Aspects and Interpretations of the Strong Vocational Interest Blank* (New York: The Psychological Corporation, 1941).

tudes help to determine the *level* at which an interest may be developed; but an individual's personality needs may be major determinants of his interests and his vocational choice.¹⁰

Stability of Vocational Interests

Strong's extensive research has shown that, although meaningful results can be obtained with able students of 14 and 15, use of his inventory with students below age 17 is not generally recommended.¹¹ Strong found an increase in stability of interests even during college years. When college graduates were retested nine or ten years after the first administration of the Strong interest inventory, the average retest correlation was .56 for those first tested as freshmen and .71 for those first tested as seniors.¹² The fact that Taylor found an average test-retest correlation of .52 for 11th-grade students retested six years later indicates fairly satisfactory stability of interest scores for senior high school students.¹³

The relative instability of vocational interests in the earlier years does not rule out the use of an inventory in the ninth or tenth grade but does suggest that greater caution should be exercised in interpreting the results. In one research study the *Kuder Preference Record* was administered in the 9th grade and again in the 12th grade.¹⁴ For only three-fourths of the students did their two highest interests in grade 9 remain among their *three* highest in grade 12. Similarly, for only three-fourths of the students did their lowest 9th-grade interest area remain among their *three* lowest in grade 12. The problem of reading difficulty also must be considered in selecting an interest inventory for use with younger students.¹⁵

TYPES OF INTEREST INVENTORIES

Test authors have used a variety of approaches in the development of interest inventories. Two of these approaches (illustrated by the Strong and Kuder inventories) are described in this chapter section.

¹⁰ J. G. Darley and Theda Hagenah, *Vocational Interest Measurement* (Minneapolis, Minn.: University of Minnesota Press), pp. 190-193.

¹¹ *Ibid.*, pp. 419-426.

¹² E. K. Strong, Jr., *Vocational Interests of Men and Women* (Stanford, Calif.: Stanford University Press, 1943), p. 363.

¹³ K. Taylor, "Reliability and Permanence of Vocational Interests of Adolescents," *Journal of Experimental Education*, vol. 10 (September 1942) pp. 81-87.

¹⁴ George G. Mallinson and William Crumbine, "An Investigation of the Stability of Interests of High School Students," *Journal of Educational Research*, vol. 45 (January 1952), pp. 369-383.

¹⁵ B. Steffire, "The Reading Difficulty of Interest Inventories," *Occupations*, vol. 26 (June 1947), pp. 95-96.

These two approaches to the measurement of interest are similar to the two major interpretations of "aptitudes" and the corresponding approaches to aptitude testing, discussed in the preceding chapter. The student will recall that some aptitude tests, for example, the prognostic tests, measured a combination of abilities that represented *aptitude for* a subject field. These tests tended to be quite heterogeneous with respect to the factors of mental ability measured. Another approach to aptitude testing was to construct homogeneous tests of *specialized aptitudes*, designed to measure discrete, unitary abilities.

In a sense, the original work on the Strong inventories is analogous to the first approach used in aptitude testing; Strong developed his interest scales on the basis of empirical data for individual items, selecting his items on the basis of their predictive validity for the criterion of *occupational membership* and making no attempt to develop homogeneous scales of unitary traits or constructs. Kuder's approach to interest measurement represents the second or unitary-trait approach. Interest dimensions were interpreted as unitary traits, and every attempt was made to increase the homogeneity of each interest scale, as is appropriate in measures of traits or constructs.

An Inventory Based on Empirical Study of Interests

Strong, who developed the first interest inventory based on extensive research, approached the problem by studying the ways in which the interests of persons successfully employed in selected occupations differed from those of men of similar age, selected at random from *occupations usually entered by college-educated men*. For each item, the percentage of men in that occupation was compared with the percentage for "men-in-general." The following example illustrates the basis for assigning weights to item 1 "actor" on the interest scale for Engineers.

GROUP	PERCENT LIKE	PERCENT INDIFFERENT	PERCENT DISLIKE
Engineers	9	31	60
Men (general group)	21	32	47
Difference	-12	- 1	13
Weight assigned	- 1	0	+ 1

On the basis of these empirical data, a "like" response to the vocation of actor is scored -1 on the engineer interest scale, while the response of dislike to this item is scored +1. An examinee's score on the engineer scale, for example, would be the sum of the weights assigned to his responses.

On the basis of his research, Strong assigned numerical weights to re-

sponses of "like," "indifferent," or "dislike" to each of the inventory items for each of his occupational scales. In each case the weights were determined by comparing the responses of men in a given occupational group (for example, actor, aviator, sales manager) with those of the combined groups of "men in general." For example, the greater the *difference* between the percentage of engineers who liked, disliked, or were indifferent to each item and the percentage of "men in general" who had the same response, the larger the weighting given that response as an indication of interest in the vocation of engineer. The weights in the revised inventory range from +4 to -4. A student who receives a high, or A, rating on the engineer scale of Strong's inventory has revealed interests similar to those reported by *engineers* in the norming sample. The letter ratings are assigned so that the top 69 percent of successful workers in the occupation (for example, engineers) would receive A, and the lowest 2 percent would receive C.¹⁶

The items in Strong's inventories include lists of (1) occupations at and above the skilled level, (2) school subjects, (3) amusements, (4) recreational and vocational activities and club offices, (5) well-known persons exemplifying occupational stereotypes or personality attributes, (6) factors affecting vocational satisfaction, and (7) self-rating questions on the present abilities and characteristics of the respondent. Different forms are provided for men and women.

Keys for at least 47 occupations are available for the SVIB for men, while the form for women can be scored for at least 27 occupations. The laborious, expensive scoring of the Strong inventories has constituted a major drawback to their use at the secondary school level. Research has demonstrated that narrowing the range of weights so that all responses are weighted as +1, 0, or -1 would result in different counseling in one twelfth to one sixth of the cases.¹⁷ Hence, Strong does not recommend such a simplification, contending that the higher validity justifies the scoring cost of approximately one dollar per student.¹⁸

As a result of factor-analysis studies of the vocational scales, several

¹⁶ Raw scores on the SVIB for each occupational criterion group were changed to *T*-scores having a mean of 50 and an *SD* of 10. Then, the *T*-scores were translated into letter grades as follows: A, *T*-scores of 45 or over; B+, 40-44; B, 35-39; B-, 30-34; C+, 25-29; C, below 24. The student can check the percentages that would fall in each group in a normal distribution; for example, approximately 69 percent of the occupational criterion group would have *T*-scores of 45 or over, corresponding to an A grade.

¹⁷ E. K. Strong, Jr., "Weighted vs. Unit Scales," *Journal of Educational Psychology*, vol. 36 (April 1945), pp. 193-216.

¹⁸ The names and addresses of organizations offering scoring services are listed in the test manual. Special answer sheets are required for scoring by the IBM electrical test-scoring machine or by the Hanks method. New scoring methods will undoubtedly reduce scoring costs.

group scales have been developed.¹⁹ These factor-analysis studies have proved valuable in classifying and interpreting the ratings received in specific occupations. When related occupations are grouped together (as they are on the report sheets supplied with the inventory), a type of *pattern analysis* is facilitated. If a student who has scored A on the physician scale (his expressed vocational choice) also has received A's or B+'s in the related occupations of group I, the counselor has a better basis for encouraging his choice (from the point of view of interest and job satisfaction) than if his ratings in these *related occupations* were largely B's and C's.

Darley makes a helpful distinction among primary, secondary, and "reject" patterns in the various occupational groups. He defines primary interest patterns as those occupational groups in which the letter ratings received by the student are predominantly A and B+; secondary patterns as those occupational groups in which B+ and B ratings predominate, and reject patterns as those fields in which the students' scores fall predominantly in the "chance-score" zone (the gray area of the test profile). He contends that it is more helpful to know that a student has primary interests in the scientific and literary occupations with a secondary pattern in social welfare than to know that he made A's as psychologist, physician, physicist, personnel director, and the like.

McArthur and Stevens,²⁰ on the basis of their 14-year follow-up study of college men, found that the scales for specific occupations had higher predictive validity than did Darley's pattern analysis. Pattern analysis, however, is usually more easily understood by students; and it is indispensable in considering vocational choices for which no occupational scale exists. Darley and Hagenah²¹ warn against using *only* the group keys in counseling students. It is quite possible that a student might have a high score on an occupation within a group and yet not have a high score on the group scale in which that occupation is classified.

Three nonoccupational scales for the SVIB have been in use for many years: the masculinity-femininity scale, an interest maturity scale, and an occupational level scale. The occupational level scale measures the similarity of one's interests to those of men in higher- or lower-status occupations. A specialization scale has been more recently developed, which

¹⁹ These clusters are enumerated in the footnote to Table 7.2. Scores on these scales do not measure degree of interest in a field but rather degree of *similarity* between the examinee's interests and those of successful workers in the occupation, or group of occupations.

²⁰ C. McArthur and Lucia B. Stevens, "The Validation of Expressed Interest as Compared with Inventoried Interests: A Fourteen-Year Follow-Up," *Journal of Applied Psychology*, vol. 39 (June 1955), pp. 184-189.

²¹ Darley and Hagenah, *op. cit.*, p. 34.

measures the extent to which one's interests resemble those of people who specialize as compared with those who might be described as generalists (for example, general practitioners in medicine).

Most of the occupational keys for the SVIB were based on the responses of men employed in the 1920s and 1930s. The need for updating one occupational key, that for psychologist, has been demonstrated. At the time that the psychologist scale was developed in 1928, Fellows of the American Psychological Association (who served as the research group) were largely experimental psychologists working on laboratory studies. Although 82 percent of the early sampling made A scores, a study made twenty years later revealed that only 52 percent of a representative group of psychologists made A scores. In fact, Kreidt²² revised the psychologist scale in terms of his findings and developed scales for specialties within psychology. A recent synthesis of research findings on this subject, however, revealed that most of the occupational fields have not shown significant changes.²³

Research has shown that women's occupational interests are not so well defined as are those of men. High correlations have been found between the housewife scale and scales for five other occupations. In fact, the "home-vs.-career decision" seems to overshadow other differences in vocational interest.²⁴

An Inventory Measuring Interests as Unitary Traits

The construction of the *Kuder Preference Record, Vocational* was based on the assumptions that (1) there are basic interest groups and (2) the student's pattern of interest can be best measured by requiring him to use the forced-choice method of requiring him to express his preferences among activities.

The method of construction of the Kuder inventories was markedly different from that used by Strong. The first step was the preparation of a large number of items that appeared to measure interest in activities in certain areas, such as literary or clerical. A lengthy preliminary edition of such items was administered to a large, unselected group of people and scored according to an *a priori* key, based on the author's judgment. Each of the items was then studied for its ability to differentiate between persons who made high and low total scores on the interest scale. Thus, a set of items was obtained in each interest field that was homogeneous or had high internal consistency.

²² P. H. Kreidt, "Vocational Interests of Psychologists," *Journal of Applied Psychology*, vol. 33 (October 1949), pp. 482-488.

²³ W. L. Layton, ed., *The Strong Vocational Interest Blank: Research and Uses* (Minneapolis, Minn.: University of Minnesota Press, 1960).

²⁴ Super and Crites, *op. cit.*, pp. 446-448.

Each item of the *Kuder Preference Record, Vocational*, is in *forced-choice* form; that is, the student is asked to choose among three activities. For example, the student indicates which of the three following choices he likes *best* and which he likes *least*.

- a. Develop new varieties of flowers.
- b. Conduct advertising campaign for florists.
- c. Take telephone orders in a florist's shop.

If the student chooses alternative *a* as the best liked, he receives credit toward his score in the scientific and artistic areas. If he prefers alternative *b*, his persuasive score is credited; if he chooses *c*, he is credited in the clerical area.

By means of a special answer pad, the inventories are easily self-scored by the students; or they can be machine scored. Scores are obtained in each of ten areas: outdoor, mechanical, computational, scientific, persuasive, artistic, literary, musical, social service, and clerical. These scores are plotted on a profile sheet for men or women students and are thereby converted into percentile ranks for those groups. Instead of comparing his responses with those of workers in various occupations, the student compares his scores in each interest area with those of "men students in general" or "women students in general."

Because of the type of item used (forced-choice or preference), Kuder scores reflect the student's *relative* interest in different fields. The "average score" for *any student* in all areas would approximate a PR of 50. The forced-choice type of item does not permit the enthusiastic student with many interests to show a higher average level of interest than does the apathetic student.

A student's profile is usually interpreted by noting his two highest interest scores and referring to a list of occupations in the manual for which this combination of scores is characteristic. Low-interest areas should also be taken into account, for the student might dislike occupations involving such activities.

This type of interpretation rests on assumptions about the relationships between interest scores and satisfaction in logically related occupations. The latest manual provides supplementary data showing profiles of average scores for various occupational groups (144 men's and 68 women's occupations). Means and *SD*'s for each occupational group are given. In general, the findings support logical expectations. For example, musicians are high in music, chemists in science, and authors in literature; there were, however, a sufficient number of exceptions to demonstrate that the validity of logical inferences should be tested empirically.

An interesting and promising device to aid in student self-appraisal and vocational guidance is the *Kuder Preference Record, Personal*, which summarizes students' preferences in such categories as: (1) preference for being active in groups, (2) preference for familiar and stable situations, (3) preference for working with ideas, (4) preference for avoiding conflict, and (5) preference for directing others. The same type of item (the forced-choice triad) is used. Since one is most interested in *differences* between scale scores, it is encouraging to find that the intercorrelations between scale scores are low. In a sense, this is a personality inventory with implications for vocational choice and job satisfaction that has the appearance of an interest inventory. A verification scale is used to identify students who answer carelessly or without understanding.

The most recent inventory in the Kuder series is the *Kuder Occupational, Form D*, published in 1956. This inventory closely resembles the Strong inventory in purpose and in techniques of construction. The one hundred forced-choice items in Form D were selected from the items in the Kuder vocational and personal inventories. This form is especially useful to companies that wish to develop keys for use in the placement of applicants in specific job situations. Until longitudinal studies provide predictive validity data, it is impossible to judge whether this inventory will prove to be as valuable in counseling as the Strong. In this test, raw scores are translated into "differentiation ratios." A positive DR for an occupational group means that a larger percentage of persons in that occupational group received the examinee's score than did the base group of "men-in-general." The higher the examinee's DR in the positive direction, the greater the probability that his interests are similar to those of workers in that occupation. The manual includes preliminary data concerning the significance of these scores in vocational counseling. Although more data are needed, findings to date are promising.²⁵

OTHER INTEREST INVENTORIES The Strong and Kuder inventories, just discussed, are the only ones on which extensive research data have been cumulated. The *Guilford-Schneidman-Zimmerman Interest Survey* has the distinction of having been based on factor analysis. Nine categories of interest have been identified, each of which has two subscores (for example, aesthetic appreciation vs. aesthetic expression). Further research studies on the relationship of student test scores to later criterion data on job success and job satisfaction would make this inventory of greater value to counselors.

Another vocational interest inventory frequently used in school counsel-

²⁵ *Ibid.*, pp. 552-560.

ing is the *Occupational Interest Inventory* by Lee and Thorpe. This inventory was devised by the methods outlined in Table 4.2, on content validity. The authors defined a universe of items, that is, the job descriptions in the *Dictionary of Occupational Titles* (a handbook prepared by the United States Employment Service). Within each of six areas (selected on logical bases rather than factor analysis), tasks were selected to represent low, medium, and high levels of responsibility. These tasks were presented in pairs as forced-choice items.

The *Occupational Interest Inventory* yields scores for six fields of interest: personal-social, natural, mechanical, business, arts, and science. In addition, responses are rescored to obtain three type-of-interest scores—in verbal, manipulative, and computational activities. A level-of-interest score, revealing whether the student has tended to choose activities at the routine, skilled, or professional-supervisory levels, is also obtained; this score is helpful to the counselor in applying inventory results to problems of vocational choice.

The OII differs from the *Kuder Preference Record* in several respects: (1) it offers the student pairs of items rather than three items for comparison; (2) all OII items are concerned exclusively with occupational activities; (3) the Kuder scales are homogeneous while those of the OII are much more heterogeneous and hence more difficult to interpret. Basing the items on the *Dictionary of Occupational Titles* may make the inventory more acceptable to adult users but may involve the inclusion of many activities with which students have little knowledge or experience.

A study of the relationship of *Occupational Interest Inventory* scores with those from the *Kuder Preference Record*²⁶ revealed the following correlations above .60:

OCCUPATIONAL INTEREST INVENTORY	KUDER PREFERENCE RECORD	<i>r</i>
Personal-social	Social service	.627
Mechanical	Mechanical	.757
Business	Clerical	.627
Sciences	Scientific	.792
Verbal	Literary	.696
Verbal	Mechanical	-.701

The correlation between the computational scores on the two inventories was only .544. Examination revealed that the computational items of the *Occupational Interest Inventory* were largely clerical, whereas the compu-

²⁶ Edward C. Roeber, "The Relationship between Parts of the Kuder Preference Record and Parts of the Lee-Thorpe Occupational Interest Inventory," *Journal of Educational Research*, vol. 42 (April 1949), p. 606.

tational scale of the Kuder included activities involving higher mathematics. Such comparisons indicate the importance of the counselor's knowing the specific content of all tests for which he interprets scores.

BASIC INTEREST GROUPS

Although various approaches have been used in the construction of interest inventories, considerable agreement has developed on the basic interest groups to be considered in vocational guidance. Table 7.2 is a version of a table prepared by Super and Crites,²⁷ modified to include the *Occupational Interest Inventory* and to exclude the Allport-Vernon²⁸ and Lurie inventories, which are infrequently used in secondary schools. This table not only provides a summary of the interest groups measured by each of the leading inventories but also includes, in the last two columns (1) the results of Guilford's latest research in this area and (2) a synthesis by Super and Crites, who present their list of basic interest groups with the justification that "a cautiously named concept, cautiously used, is better than no concept at all."²⁹

VALIDITY OF INTEREST INVENTORIES

Prediction of Success in School Subjects

Interest inventory scores have seldom shown correlations above .30 with grades or achievement test results in specific subject fields.³⁰ One study indicated that interest scores did predict *differences* in achievement between courses; for example, engineer interests on the *Strong Vocational Interest Blank* correlated .61 with the difference between grades in mathematics and history.³¹ Although one must avoid generalizing from a single study, this finding is in accord with the common-sense assumption that

²⁷ Super and Crites, *op. cit.*, p. 382.

²⁸ A revision of the well-known Allport-Vernon scale was published in 1951. Extensive research on this and the earlier edition is available. Its use is ordinarily limited to college students and adults. However, the inventory can appropriately be used with superior high-school students—for example, in elective courses in psychology (G. W. Allport, P. E. Vernon, and G. Lindzey, *A Study of Values*, rev. ed. (New York: The Psychological Corporation, 1951)).

²⁹ Super and Crites, *op. cit.*, p. 383.

³⁰ *Ibid.*, pp. 433-435.

³¹ David Segel, "Differential Prediction of Scholastic Success," *School and Society*, vol. 39 (January 20, 1934), pp. 91-96.

Table 7.2
Basic Interest Groups as Defined in Five Interest Inventories,
and as proposed by Super and Crites^b

THURSTONE	STRONG	KUDER	O.I.I.	GUILFORD ^a	SUPER AND CRITES ^b
Science	Science (Groups I, II) ^c	Scientific	Science	Scientific	Scientific
People	People (Group V)	Social service	Personal Social	Social Welfare	Social Welfare
Language	Language (Group X)	Literary	(Verbal) ⁴		Literary
	Things vs. people (Group IV)	Mechanical	Mechanical	Mechanical	Material (concrete)
Business	Business detail (Groups VII and VIII)	Clerical; Computa- tion	Business computa- tional ^d	Clerical	Systematic (record- keeping)
	Business contact (Group IX)	Persuasive		Business	Contact (with people for material gain)
	Musician (Group VI)	Musical	Arts	Aesthetic expression	Aesthetic expression
		Artistic		Aesthetic interpreta- tion	Aesthetic interpreta- tion
		Outdoor	Natural	Outdoor	

^a J. P. Guilford, and others, "A Factor Analysis of Human Interests," *Psychological Monographs*, No. 375 (Washington, D.C.: The American Psychological Association, 1954).

^b After a study of the factors appearing in the Thurstone, Allport-Vernon, Lurie, Strong, and Kuder tests, together with the literature upon which they are based, Super and Crites developed

Table 7.2 (Continued)

Basic Interest Groups as Defined in Five Interest Inventories,
and as proposed by Super and Crites^b

the list of factors appearing in the last column of this table. Donald E. Super and John O. Crites, *Appraising Vocational Fitness by Means of Psychological Tests* (New York: Harper & Row, Publishers, Inc., 1962), pp. 383-384.

^c A factor analysis by Strong of the *Vocational Interest Blank for Men* (revised) made at a time when only 36 occupational scales were available led to the formulation of the following groups: Group I—artist, psychologist, architect, physician, dentist; Group II—mathematician, physicist, engineer, chemist; Group III—production manager; Group IV—aviator, farmer, carpenter, painter, mathematics-science teacher, policeman, forest service; Group V—YMCA physical director, personnel manager, YMCA secretary, social science teacher, school superintendent, minister; Group VI—musician; Group VII—certified public accountant; Group VIII—accountant, office worker, purchasing agent, banker; Group IX—sales manager, real estate salesman, life insurance salesman; Group X—advertising man, lawyer, author-journalist; Group XI—president of manufacturing corporation. Group scales are available for only Groups I, II, V, VIII, IX, and X.

^d A summary score on type of interest, obtained by a regrouping of items already included in scoring for the six fields of interest.

interest would usually affect the channeling of effort, especially by noncompulsive students.

Prediction of Occupational Choice and Job Satisfaction

A number of research studies have indicated that persons who choose occupations consistent with their interest inventory scores tend to be more satisfied with their jobs than those who do not.³² Sarbin and Anderson³³ found that 82 percent of their men clients who expressed dissatisfaction with their work were engaged in occupations that were inconsistent with their primary interests.

Strong's follow-up studies indicated that men who remain in an occupational field score significantly higher in that field than men who change to another field. Moreover, men who change from one vocation to another tend to change into a field in which their interest scores are higher than for their earlier choice. Impressive predictive validity data have come from

³² Laurence Lipsett and James W. Wilson, "Do 'Suitable' Interests and Mental Ability Lead to Job Satisfaction?" *Educational and Psychological Measurement*, vol. 14 (Summer 1954), pp. 373-380. Dallas K. Perry, "Validities of Three Interest Keys for United States Navy Yeomen," *Journal of Applied Psychology*, vol. 39 (April 1955), pp. 134-138.

³³ T. R. Sarbin and H. C. Anderson, "Preliminary Study of the Relation of Measured Interest Patterns and Occupational Dissatisfactions," *Educational and Psychological Measurement*, vol. 2 (January 1942), pp. 23-36.

Strong's follow-up studies of the later occupational status of men who had taken the SVIB when they were college students. The higher a student's standard score in a scale, the greater the probability that he would enter the occupation suggested by that scale. Typically, the occupation in which a student was working ten or more years later ranked second and third for him among all the occupational scales of the SVIB.³⁴

McCully³⁵ did a similar follow-up study on men who had taken the Kuder in Veterans Administration counseling centers several years before. He found that persons engaged in accounting and related fields had had unusually high scores in the computational and clerical areas; those employed in engineering and related occupations had scored relatively high in scientific and mechanical; those engaged in high-level sales work had averaged high in the persuasive area; while those in mechanical repairing, electrical repairing, and bench crafts had scored highest in the mechanical area.

Most of the studies on the predictive validity of interest inventories have been confined to students who enter the professions or skilled trades. Interest inventories probably have little predictive value for occupational choice among semiskilled and unskilled jobs. Studies of job satisfaction have indicated that only people from the higher socioeconomic levels tend to mention interest in their work as a source of job satisfaction. Those at the lower levels stress job security, economic returns, and recognition as a person.³⁶ When Clark³⁷ used Strong's approach in an attempt to develop interest keys for various trades, he found differential interest patterns within the skilled trades but not among the unskilled occupations.

Prediction of Success in Vocations and Vocational Training Courses

Disappointing results have been obtained with the use of interest inventory scores as predictors of *success* in vocational training. Air Force studies³⁸ revealed that almost all *r*'s between interest scores and grades in thirteen vocational training schools were below .20. Interest scores, however, do have predictive validity when the criterion is *completion* of the vocational training.³⁹

³⁴ E. K. Strong, *Vocational Interests of Men and Women* (Stanford, Calif.: Stanford University Press, 1943).

³⁵ C. Harold McCully, "The Validity of the Kuder Preference Record." Unpublished doctoral dissertation, George Washington University, 1954.

³⁶ Darley and Hagenah, *op. cit.*, pp. 8-10.

³⁷ Kenneth E. Clark, *The Vocational Interests of Non-Professional Men*. (Minneapolis, Minn.: University of Minnesota Press), 1961.

³⁸ W. L. Layton, *The Strong Vocational Interest Blank: Research and Uses*. (Minneapolis: University of Minnesota Press, 1960).

³⁹ Strong, 1943, *op. cit.*, p. 524.

In order to construct a vocational interest scale that had a fairly high correlation with *success* in a specific vocational field, one would have to select items *that differentiated between successful and unsuccessful men within that vocational field*. The rationale for constructing prediction tests is summarized in Table 4.5. SVIB scales, for example, were devised to differentiate *between* men in an occupation and professional men in general, rather than to differentiate between the successful and unsuccessful *within* an occupation. In other words, if the criterion of success in an occupation (rather than membership in an occupation) had been used as the basis for selecting items, the tests' predictive validity for success criteria would undoubtedly have been increased. Such a procedure would be desirable if an interest inventory were to be used in selection.

It seems logical that interest in one's field of work would be highly correlated with success in those occupations in which success depends largely on one's commitment and enthusiasm. In one such vocation, insurance sales, Strong found that 56 percent of the men with A scores, as compared with only 6 percent of those with C scores, sold sufficient insurance to make an adequate income.⁴⁰ In a study of newly employed insurance salesmen, Bills⁴¹ found a 76 percent failure rate among those with low scores on the life insurance salesman and real estate salesman scales; among those with high scores on these scales, the failure rate was only 22 percent. Other research studies have revealed positive relationships between relevant SVIB scores and success in engineering, advertising, and technical foremanship.⁴² In the well-known follow-up study of gifted individuals by Terman and Oden,⁴³ those rated as "least successful" included five times as many men with low interest scores in their current occupations than did the "most successful" group.

A promising approach to the study of the predictive validity of interest inventories was used by Frederiksen and Melville,⁴⁴ who studied the relationship of interest and ability test scores to achievement in engineering. They found that interest in engineering showed a negligible relationship to achievement for compulsive students,⁴⁵ who work hard on tasks

⁴⁰ *Ibid.*, pp. 487-488.

⁴¹ M. A. Bills, "Relation of Strong's Interest Blank to Success in Selling Casualty Insurance," *Journal of Applied Psychology*, vol. 22 (December 1938), pp. 97-104.

⁴² *Ibid.*, pp. 501-504.

⁴³ L. M. Terman and M. H. Oden, *The Gifted Child Grows Up* (Stanford, Calif.: Stanford University Press, 1948).

⁴⁴ Norman Frederiksen and Donald S. Melville, "Differential Predictability in the Use of Test Scores," *Educational and Psychological Measurement*, vol. 14 (Autumn 1954), pp. 647-656.

⁴⁵ The researchers used two indicators of compulsiveness: (1) having interests like those of professional accountants and (2) having relatively low speed of reading scores, in comparison with vocabulary scores on a standardized reading test.

whether they are interested or not. However, for noncompulsive students, who work hard only when interested, interest scores were significantly correlated with achievement in engineering. While it is always unwise to generalize from a single research study, it seems reasonable that a person with adequate abilities but low interest *can* do well in a field but *may* not if he is noncompulsive.

INTERPRETATION OF INTEREST-INVENTORY RESULTS

In Chapter 17, an illustration is given of the use of interest-inventory scores in counseling, in combination with aptitude test results and other relevant data. This section is concerned chiefly with the problems of interpretation that grow out of the limitations of interest inventories and the inadequacy of research data needed as a basis for valid inferences from test scores.

The trained counselor recognizes the comparatively low relationships between interests and abilities, as well as the critical importance of considering *both* ability and interest in evaluating vocational choices. In many school situations, however, various factors have led to an enthusiastic, uncritical use of interest inventories. Since interest-inventory results are not threatening to students, the inventories are often self-scored and the results graphed by students in a group situation. Although this process can be a helpful one, there is real danger of misinterpretation when interest profiles are analyzed by students without personal assistance from a well-trained guidance worker and in the absence of data about aptitudes and other limiting factors. Procedures are outlined in Chapter 17 to minimize the dangers of such misinterpretation.

One need only read through an interest inventory to realize that many responses of a junior high school student may be invalid because of his lack of experience. The adolescent's vocational interests are part of his changing, gradually emerging concept of self; they tend to become more stable and realistic as he matures. These factors do not rule out the use of interest inventories in junior high school as part of the process of helping young people in self-appraisal and stimulating them to consider a wider variety of vocational opportunities. They do, however, suggest caution in the interpretation of inventory results as indicators of suitable vocational choices. Many students at the junior high school level do not have well-defined patterns of interest that have implications for vocational choice. For such students, of course, interest-inventory results are of little value.

The need for more adequate research data on the significance of interest-inventory results cannot be overemphasized. Until such time as data are available on the interest patterns of skilled and semiskilled workers,

interest-inventory results will be of limited value for a large proportion of secondary school students.⁴⁶

More studies on the interrelationships among scores on the various inventories would be helpful to counselors. Counseling experience has indicated, for example, that apparent discrepancies between Kuder and Strong scores may be significant for vocational guidance.⁴⁷ For example, some persons who had high persuasive scores on the Kuder inventory but low life-insurance-salesman scores on the Strong seemed, on the basis of interview and case-history material, to be interested in promotional activities but to dislike activities in which they need to push people to the point of action, as in closing a sale. Research data supporting such hypotheses as this would help counselors to do a more professional job in the interpretation and use of test data.

Although job applicants can and do fake results on interest inventories, it would seem that students seeking guidance would report frankly concerning their likes and dislikes. Even under these circumstances, however, a student's objectivity may be reduced by his desire to see himself in a preferred occupational role and to choose activities that are associated with that role in an occupational stereotype. When students are instructed to answer the Strong inventory so as to obtain high scores in a specified occupation, the majority do receive A ratings. Research findings indicate that the Strong inventory is more susceptible to upward faking, and the Kuder to downward faking.⁴⁸ Through the use of verification scales, it is possible to identify examinees who try to fake, as well as those who succeed.

Despite their limitations, interest inventories can be very helpful in aiding students to focus attention on one or two fields of choice, in stimulating them to study possible vocations, in suggesting occupations not previously considered and fields in which special aptitude tests may be desirable, in helping girls to clarify their thinking in the career-vs.-home decision, and in helping students to distinguish between genuine interests and expressed interests based on extrinsic pressures (parental aspirations, hero worship, and the like).

⁴⁶ Dr. Kenneth Clark has been systematically studying the problems involved in the measurement of interests at the lower occupational levels. Several years of research with enlisted men, conducted for the Office of Naval Research, has resulted in the development of an unpublished interest inventory, the *Minnesota Vocational Interest Inventory*. When this inventory is published, it should help in the counseling of students who are choosing among such skilled and semiskilled occupations as retail sales work, truck driving, and working in the various building trades.

⁴⁷ Donald E. Super, "The Kuder Preference Record in Vocational Diagnosis," *Journal of Consulting Psychology*, vol. 11 (July-August 1947), pp. 184-193.

⁴⁸ H. P. Longstaff, "Fakability of the Strong Interest Blank and the Kuder Preference Record," *Journal of Applied Psychology*, vol. 32 (August 1948), pp. 360-369.

An invitation to discuss the results of an interest inventory often brings to the counselor's office students who need help on a variety of significant problems. In the course of his discussion regarding vocations, the student will naturally discuss any academic difficulties he may be having and will frequently bring in problems in his relationships with friends, family, or employer that he needs to discuss with his counselor.

MEASUREMENT OF ATTITUDES

Definitions of Terms

There is no clear-cut distinction between (1) questionnaires regarding an examinee's likes, dislikes, and preferences, which are called *interest inventories*, and (2) questionnaires regarding his attitudes, which are typically organized into *attitude scales*. Greene⁴⁹ makes the distinction that interest inventories are usually concerned with a person's feelings or preferences regarding personal activities, and the choices involved have no moral connotations; while attitude scales assess the person's position on a continuum of approval-disapproval toward social institutions, group activities, and principles that affect the welfare of others. In this sense of the word, a person with a high interest in an occupation, such as engineering, is one who personally enjoys many of its activities, while a person who has a favorable attitude toward engineering may be one who values it as a significant, high-status occupation even though he might personally dislike the activities it involves.

Attitudes may be defined as "predispositions to react negatively or positively in some degree toward an object, institution, or class of persons."⁵⁰ For example, there are many ways in which a person might react negatively or positively toward the profession of engineering. A person with a high positive attitude toward the profession might encourage students to enter it, might listen respectfully to the views of engineers on civic problems, or might vote for a candidate for local office largely because he was an engineer.

Various Approaches to the Study of Attitudes

OBSERVING MANIFEST ATTITUDES Obviously, one approach to studying attitudes is to observe them as they are *manifested* in behavior. For ex-

⁴⁹ Edward B. Greene, *Measurements of Human Behavior* (New York: The Odyssey Press, 1952, p. 594).

⁵⁰ Jum C. Nunnally, Jr., *Tests and Measurements: Assessment and Prediction*. (New York: McGraw-Hill Book Company, Inc., 1959), p. 300.

ample, science teachers who wish to develop in students the "scientific attitude" (a strong positive attitude toward the methods of science), could observe evidences of such an attitude in students' laboratory work, group discussions, and other aspects of science study. Although it is doubtful that the teacher would obtain a sufficiently large sample of behaviors to justify reliable appraisal of this attitude in individual students, sufficient evidence might be obtained to justify conclusions regarding group progress.

In an attempt to appraise behavior changes in students as a result of science instruction, West developed several behavioral descriptions to be used by trained observers during instruction periods in science. The descriptions for *critical-mindedness* and *open-mindedness* are quoted below:

Critical-mindedness. Enter the code number *1b* against the name of each pupil for evidences of critical-mindedness toward the class situation as shown by weighing evidence with respect to its pertinence, soundness, or adequacy. Examples are: Pupil asking for statement of source of information before accepting it; pupil verifying statements read or heard; pupil questioning authority constructively; pupil questioning truth of statement before he is willing to accept it as final.

Open-mindedness. Enter the code number *1c* against the name of each pupil for evidence of being willing to abandon predetermined ideas in favor of ideas which seem to be more nearly correct. Examples are: Pupil accepting good, clear evidence without useless argument; pupil welcoming suggestions and information about new undertakings; pupil accepting evidence which modifies false beliefs; pupil willing to respect the viewpoint of others; pupil willing to acknowledge his ignorance of the situation; pupil exhibiting conduct which shows that he is free from prejudices; pupil taking criticism kindly and attempting to profit by it.⁵¹

Teacher effectiveness in the observation of attitudes, as manifested in behavior, may be heightened by taking advantage of special situations in which student attitudes are more readily revealed and observed. Students at work in a laboratory situation or discussing the results of a teacher demonstration are likely to offer evidences of their open-mindedness, habit of suspending judgment, habit of looking for natural causes, and other scientific attitudes. Similarly, attitudes toward civic participation might be manifested as students work together in student-government activities.

The teacher can set up special situations for the observation of attitudes, for example, asking students to act out possible solutions to unfinished

⁵¹ J. Y. West, *A Technique for Appraising Certain Observable Behavior of Children in Science in Elementary School* (New York: Bureau of Publications, Teachers College, Columbia University, 1937), p. 16.

stories⁵² or films, or assigning students to committee work on problems that require teamwork for their solution.

STUDYING EXPRESSED ATTITUDES The preceding examples have involved *manifest* attitudes, as evidenced in student behavior. The approaches were similar to those used in the study of manifest interests, that is, observing and recording relevant behaviors of the student. Two other approaches, each paralleling an approach used in the study of interests, are (1) to obtain direct *expressions* of attitudes, and (2) to develop questionnaires or *inventories* of attitudes, in which the examinee reacts to a large sampling of verbal statements. Just as in the study of interests, we find that the latter approach (because of the larger sampling and the minimizing of the effect of stereotypes) tends to yield more reliable results than requesting direct expressions of over-all attitudes toward an institution or class of people. That is, summing students' reactions to a series of selected statements regarding church, war, or a specific ethnic group usually gives more consistent results (from one time to another) than requesting an over-all expression of attitude.

Attitude inventories are similar to interest inventories in at least two ways: (1) they are self-report questionnaires in which the examinee can usually modify his responses if he desires to do so; (2) they contain a sampling of verbal statements to which the examinee gives responses. However, attitude inventories differ from interest inventories, in the way in which items are selected and scored. The items on attitude inventories are selected to *represent* a universe of attitudes toward the object, institution, or class of persons. They are *not* selected for their correlation with some external criterion. They are used for the first and fourth purposes, rather than the second and third purposes, outlined in the discussion of validity in Chapter 4. Hence, their content and construct validity are usually more important than their concurrent or predictive validity.⁵³

In social studies classes, or any other situations in which a variety of defensible attitudes might be held, the teacher must *not* grade or otherwise evaluate students on the basis of attitudes they hold. The teacher's legitimate concern is with the student's ability to formulate and defend his own attitudes with respect to labor-management or other social studies problems, rather than with the specific attitudes held. When we study

⁵² See George Shafel and Fannie R. Shafel, *Role Playing the Problem Story* (New York: National Conference of Christians and Jews, 1952). The use of reaction stories is discussed further in Chapter 8 of this textbook.

⁵³ An exception would be an inventory of study attitudes, such as the *Survey of Study Habits and Attitudes* by Brown and Holtzman, which would be expected to have predictive validity for the prediction of grade-point average in academic work. Publication data for this and other inventories are given in Appendix A.

attitudes in an area in which competent judges would defend different points of view, we must describe attitudes rather than evaluate them. If the teacher imposes a right vs. wrong type of evaluation, in such an area "expected answers" will be given; and the entire procedure becomes indefensible.

Since appraisal of *individual* attitudes as a basis for grading is inappropriate, published attitude questionnaires have been used chiefly in research work. They may also be appropriately used for checking on shifts of attitude in classes or groups as a basis for evaluating the effectiveness of instruction, *provided* that they are administered anonymously. For example, a teacher might wish to know how the group attitude had changed toward some health practice, some foreign country, or the United Nations, as the result of a unit of work that included reading and discussion concerning any of these areas.

If student responses to an attitude questionnaire can be summarized in such a way as to locate each examinee at some point on a continuum from "strongly negative attitude" through "neutral" to "strongly positive attitude," the questionnaire can be called an attitude scale. To develop an attitude scale, a series of statements representing different degrees of positive and negative attitudes must be formulated. In order to achieve content validity, the statements might well be based on the results of interviewing a representative sampling of subjects or analyzing statements of attitude written by such a representative group.

THE THURSTONE METHOD OF ATTITUDE-SCALE CONSTRUCTION An early method of constructing attitude scales, which is still widely used, was developed by Thurstone and his associates.⁵⁴ After a large number of statements (which expressed various degrees of positive and negative feeling toward some institution or group) are obtained, each statement is reproduced on a card or slip of paper. Then, a large number of judges independently sorted these slips according to their position on an 11-point continuum (ranging from "extremely favorable" through "neutral" to "extremely unfavorable"). The judges do not give their personal reactions to the statements but arrange them on a continuum of intensity of positive and negative feeling.

Items that are assigned a variety of values by the judges are eliminated as ambiguous or as unrelated to the attitude being judged. Only those statements that show relatively low interjudge variability are retained. From among these statements, there would be selected 15-25 statements that were fairly evenly spaced, with respect to median rating, on the

⁵⁴ L. L. Thurstone and E. J. Chave, *The Measurement of Attitude* (Chicago: University of Chicago Press, 1929).

attitude continuum. For example, on an 11-point scale, items might be selected with median intensity values of 0, 0.5, 1.0, 1.5, and the like, through 11.

Once the statements have been selected, they are arranged in *random* order on the printed form. The student then marks those statements with which he agrees; and his score is the median intensity value of statements he has marked. Scales of the Thurstone type have been constructed for attitudes toward church, war, censorship, capital punishment, and many other institutions and issues, as well as for attitudes toward number of ethnic groups. The scales are quickly administered and easily scored. The method of scaling is objective and reliable because many independent judgments are used.

THE LIKERT METHOD OF ATTITUDE-SCALE CONSTRUCTION The first step in the Likert method is also the collection of a large number of statements expressing various degrees of positive and negative feelings about an object, institution, or class of persons. The selection of items for the attitude scale, however, does not involve the use of judges; rather, the selection is based on the results of administering the items to a representative group of subjects. Each item is rated by subjects taking the attitude scale on a five-point continuum from "strongly approve" to "strongly disapprove." The total score is the sum of all the item scores.⁵⁵ The validity of each item is studied, with the criterion being total score on the preliminary edition. Only those items that have high correlations with total score are retained. Items that have low correlations are excluded as either unreliable or as measuring some extraneous attitude factor. As a result, the shorter revised attitude scale is much more homogeneous than the preliminary edition. It has greater internal consistency, a characteristic which is necessary, but not sufficient, for construct validity.

The advantages of the Likert method include (1) greater ease of preparation; (2) the fact that the method is based entirely on empirical data regarding subjects' responses rather than subjective opinions of judges; (3) the fact that this method produces more homogeneous scales and increases the probability that a unitary attitude is being measured; and (4) the scales provide more information about the subject's attitudes, since an intensity reaction is given to each of many items. The chief disadvantage of the Likert method is that the scores are relative to the group used in scale construction, whereas the Thurstone method establishes a neutral point as a basis of reference. If Likert-type scales were normed

⁵⁵ Each favorable item is scored by giving five points for "strongly agree," four points for "agree," three for "uncertain," and the like. The scoring is reversed for unfavorable statements.

on representative groups, the scores could be more adequately interpreted. It is doubtful, however, that such norming will be done. Attitude scales are used chiefly in research studies, and most researchers find that an attitude scale devised to suit their specific purpose is more suitable than any of the published scales.⁵⁶

The Thurstone and Likert-type scales for the same institution or group tend to yield results that agree or intercorrelate highly. Reliability coefficients for such scales tend to be in the .80's and are highly satisfactory for group comparisons. Recent studies on attitude-scale construction evidence a trend toward the development of homogeneous subscales *within* a larger area, for example, toward aspects of an institution such as the church or the United Nations.

The chief criticism that might be leveled at attitude scales is concerned with the indirectness of measurement, that is, verbal statements are used as a basis for inferences about "real attitudes." Moreover, attitude scales are easily faked. Although administering the scales anonymously may increase the validity of results, anonymity makes it difficult to correlate the findings with related data about the individuals unless such data are obtained at the same time. It seems that we must limit our inferences from attitude-scale scores, recognizing that such scores merely summarize the verbalized attitudes that the subjects are willing to express in a specific test situation. The student will recognize the difficulty of studying the concurrent validity of verbal attitude scales by studying their relationship with behavioral criteria. Because of many factors, persons with the same attitudes will manifest different behaviors.

SUMMARY STATEMENT

Four types of data concerning the interests of students can be obtained: (1) expressed (verbal) interest in specific occupations or activities; (2) manifest interest, as evidenced by the student's actual participation in some activity; (3) tested interest, as reflected in the student's informational background and vocabulary in special interest fields; and (4) inventoried interest, as summarized in a pattern of scores indicating the student's preferences for types or areas of activity. Almost all interest measures developed and standardized for school use are of the fourth type.

Although interests and abilities tend to be related, students must not be advised to enter occupations in which they show a high interest level *unless* data indicate that the students also have the abilities requisite for success.

⁵⁶ The research worker who wishes to develop his own scales will find Edwards' manual a valuable source of information. This manual illustrates step-by-step procedures for the Thurstone and Likert approaches, as well as the more recently developed Guttman approach. Allen L. Edwards, *Techniques of Attitude Scale Construction* (New York: The Psychological Corporation, 1957).

Research data concerning the stability of students' vocational interests indicate that the interests of young adults (that is, those in the upper college years) are much more stable than those of younger students. For this reason, Strong does not recommend the use of his inventories with students below age 17. The relative instability of the vocational interests of younger students does not rule out the use of interest inventories in the ninth or tenth grades, but it does imply the need for greater caution in the interpretation and use of results.

Two very different approaches have been used in the development of the leading interest inventories. A student's scores on the Strong inventories indicate the extent to which his expressions of like or dislike for specific activities, school subjects, and so on are similar to those of employed workers in each of more than thirty occupational groups. That is, a student with an A rating on the Actor scale has interests that agree much more closely with those of actors than with those of "men in general." The *Kuder Preference Record* requires the student to indicate his preferences within groups of three activities; such indications of preference are then summarized according to the interest areas (clerical, artistic, and the like) which they represent.

Although different approaches have been used in the construction of interest inventories, an examination of five of the leading measures (Table 7.2) reveals considerable agreement on the basic interest groups. On the basis of their analysis, Super and Crites proposed the following eight basic interest groups: scientific, social welfare, literary, material (concrete), systematic (record-keeping), contact, aesthetic expression, and aesthetic interpretation.

As a basis for more discriminating use of interest-inventory results in counseling, a number of cautions concerning the interpretation of interest scores were formulated: (1) Interest inventory results must not be interpreted as indicators of occupations in which students will be successful. (2) The relative instability of vocational interests among younger students suggests the need for great caution in the interpretation of interest scores for junior high school students. (3) A large number of students do not have well-defined patterns of interest that have implications for vocational choice. (4) Until further research reveals more well-defined interest patterns for semiskilled workers and for workers in the leading occupations for women, interest-inventory results will be of limited value for a large percentage of high school students.

Techniques of studying attitudes manifested in student behavior were briefly reviewed, as well as two leading methods of attitude-scale construction.

SELECTED REFERENCES

- BERDIE, RALPH F., "Strong Vocational Interest Blank Scores of High School Seniors and Their Later Occupational Entry," *Journal of Applied Psychology*, vol. 44 (June 1960), pp. 161-165.
- CRAVEN, E. C., *The Use of Interest Inventories in Counseling*. Professional Guidance Series. Chicago: Science Research Associates, 1961.
- DARLEY, J. G., *Clinical Aspects and Interpretation of the Strong Vocational Interest Blank*. New York: The Psychological Corporation, 1941.
- , AND THEDA HAGENAH, *Vocational Interest Measurement: Theory and Practice*. Minneapolis, Minn.: University of Minnesota Press, 1955.

- DURNALL, EDWARD J., JR., "Falsification of Interest Patterns on the Kuder Preference Record," *Journal of Educational Psychology*, vol. 45 (April 1954), pp. 240-243.
- GUILFORD, J. P., AND OTHERS, "A Factor Analysis Study of Human Interests," *Psychological Monographs*, vol. 68, No. 4. Washington, D. C.: American Psychological Association, 1954.
- KATZ, MARTIN R., "Interpreting Kuder Preference Record-Vocational Scores: Ipsative or Normative?" *The Vocational Guidance Quarterly*, vol. 10 (Winter 1962), pp. 96-100.
- KUDER, G. FREDERIC, AND BLANCHE B. PAULSON, *Exploring Children's Interests*. Chicago: Science Research Associates, 1951.
- LAYTON, WILBUR L., *Counseling Use of the Strong Vocational Interest Blank*. Minnesota Studies in Student Personnel Work No. 9. Minneapolis, Minn.: University of Minnesota Press, 1958.
- , *The Strong Vocational Interest Blank: Research and Uses*. Minneapolis, Minn.: University of Minnesota Press, 1960.
- LONGSTAFF, H. P., "Fakability of the Strong Interest Blank and the Kuder Preference Record," *Journal of Applied Psychology*, vol. 32 (August 1948), pp. 360-369.
- MCARTHUR, CHARLES, "Long-Term Validity of the Strong Interest Test in Two Subcultures," *Journal of Applied Psychology*, vol. 38 (October 1954), pp. 346-353.
- , AND LUCIA BETH STEVENS, "The Validation of Expressed Interests as Compared with Inventoried Interests: a Fourteen Year Follow-Up," *Journal of Applied Psychology*, vol. 39 (June 1955), pp. 184-189.
- MALLINSON, GEORGE G., AND WILLIAM M. CRUMBINE, "An Investigation of the Stability of Interests of High School Students," *Journal of Educational Research*, vol. 45 (January 1952), pp. 369-383.
- REMMERS, H. H., *Introduction to Opinion and Attitude Measurement*. New York: Harper & Row, Publishers, Inc., 1955.
- STEPHENSON, RICHARD R., "A New Pattern Analysis Technique for the SVIB," *Journal of Counseling Psychology*, vol. 8 (Winter 1961), pp. 355-362.
- STRONG, EDWARD K., JR., *Vocational Interests 18 Years After College*. Minneapolis, Minn.: University of Minnesota Press, 1955.
- , *Vocational Interests of Men and Women*. Stanford, Calif.: Stanford University Press, 1943.
- SUPER, DONALD E., AND JOHN O. CRITES, *Appraising Vocational Fitness by Means of Psychological Tests*. New York: Harper & Row, Publishers, Inc., 1962, Chapters 16-18.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. Outline a talk to be presented to a group guidance class before students take and self-score an interest inventory. Emphasize the values and limitations of a specific interest inventory, other sources of data on interests, the relationship between interests and aptitudes, and the like.
2. Obtain interest-inventory data for two or three high school students. Draw the test profiles for each student. Record any data given on the cumulative

record regarding expressed interests (recorded vocational choices) and manifest interests (choice of elective subjects, participation in extracurricular activities, and the like). Summarize all data on interests and indicate three or four occupations suggested by the interest pattern and general intelligence level of each student. What aptitude tests would you like to administer to each of these students to obtain necessary additional information?

3. What problems are involved in the use of interest inventories in an employment situation? Examine a specific interest inventory to see how applicants could intentionally misrepresent their pattern of interests. Consult the *Buros' Mental Measurements Yearbook*, as well as research studies on the extent to which responses can be faked on this inventory.

4. Compare the methods used in constructing and validating the Strong and the Kuder interest inventories.

5. Of what value are level-of-interest scores, such as those obtained on the Occupational Interest Inventory.

6. What are the advantages and disadvantages of using the names of occupations as items in an interest inventory?

7. What informal methods could you use to obtain evidence as a basis for judging students' interests in your subject field?

8. What are some of the problems involved in the testing of attitudes? What procedures seem to yield the best results?

9. Observe students in a discussion group and make anecdotal recordings concerning the social attitudes evidenced in their discussion.

10. Prepare a list of behavioral evidences of the scientific attitude in the subject area in which you will teach.

Informal Methods of Studying Personal-Social Adjustment

As the reader studies this and the succeeding chapter, it will become increasingly evident that he should not expect to have as much success in evaluating students' personal-social adjustment as in evaluating their scholastic aptitude or their skills in reading and arithmetic. Not only are the techniques of evaluation less well developed, but the most valid techniques are available only to the clinical psychologist or the psychiatrist. Effective use of personality inventories and projective techniques (discussed in Chapter 9) demands more training, more time, and a greater insight into individual cases of maladjustment than the classroom teacher can be expected to achieve.

Why, then, should teachers and counselors attempt evaluation in this difficult area? The answer lies largely in the importance of the child's personal-social adjustment to every aspect of his development. The student's mental health affects his ability to learn, his interest in learning, his ability to contribute to classroom experiences, and his later success as a citizen, an employee, and a parent. Attempts to help the student who is retarded in reading or any other area of school work often reveal that the student's learning difficulties are inextricably related to his problems of personal-social adjustment.

Teachers are able to observe a student's reactions in a wide variety of situations that reveal his feelings and attitudes—situations that involve social participation, rivalry, outside authority, success and failure, praise and blame. They are in a unique position to identify and refer to guidance specialists those students who are unusually tense or withdrawn, or show other evidences of poor mental health.

The results of any one technique of evaluating student adjustment may have low reliability and validity. The teacher, however, has an opportunity to study many samples of student behavior and to note their con-

sistency; to make hypotheses concerning possible causal factors contributing to poor adjustment; and then to examine these hypotheses in the light of additional information.

THE NATURE OF PERSONAL-SOCIAL ADJUSTMENT

Teachers and counselors want to contribute to the mental health of students, rather than to help them develop any specified pattern of personality traits. Defining good mental health, or good personal-social adjustment, however, is not as simple as the defining of many other educational goals. We shall begin with a brief examination of three major concepts in this area. These concepts of *maturity*, *normality*, and *adjustment* will be found to overlap. Each of them, however, contributes to an understanding of the nature of personal-social adjustment.

Maturity

The concept of maturity is especially significant to those who work with growing children. Every teacher has students in his classes who may best be described as relatively immature in certain respects, whose behavior is characteristic of a younger age group. A second-grade child who continually interrupts the teacher, a fourth-grader who tattles, or a junior high school student who frequently complains to the teacher about other students is not functioning at the expected level of maturity. Buhler and others use the term "average developmental expectation" for such "age norms." They point out that teachers have difficulty with students who have not developed "school maturity" in the sense of willingness and ability to carry out work assignments at the sacrifice of immediate impulse satisfactions.¹

The term "average developmental expectation" implies that a student's maturity should be judged in comparison with others of *his own age and culture group*. Children cannot be judged in terms of their progress toward adult behavior. Docile children can learn to display behavior typical of much older children or of adults, but such docility should not be mistaken for genuine maturity. In fact, in the preadolescent or adolescent years, such conforming behavior may indicate immaturity and prolonged dependence upon adult approval.

A child whose behavior is fairly typical of children of his age is considered to be at a satisfactory level of maturity; a child whose behavior

¹ Charlotte Buhler, Faith Smither, and Sybil Richardson, *Childhood Problems and the Teacher* (New York: Holt, Rinehart and Winston, Inc., 1952), p. 43.

is like that of much younger children is considered immature. It is difficult to describe an above-average degree of maturity in children without including in the "superior" group those whose pseudomaturity may be achieved at the cost of internal conflict as well as less social acceptance within their own age group.

Normality

In one sense of the word, "normal" implies average or typical. Statisticians use the word in that sense. In many aspects of personal-social adjustment, behavior is perhaps best described as being similar to, or deviating from, average behavior. It is normal, for example, for children to become restless after a long period of drill work or for adolescents to show less respect toward adults than do younger children. No value judgment is implied in this use of the term "normal."

Deviations from average behavior are signals to the teacher that a student should be observed more carefully. If a student is unusually tired after physical exercise, unusually tense when called upon to recite, unusually excited when he is given a small role in the class play, or unusually depressed when he is criticized, the teacher will want to discover the reasons behind this deviating behavior.

It is normal for children to have problems. When an individual is unable to cope successfully with his problems, he may show abnormal behavior—not only in a statistical sense but in the sense of its being inappropriate, ineffectual, unwholesome, or self-defeating. In this sense of the word, "normality" is used with the meaning of "health" vs. "illness." "Abnormality" implies *malfunctioning* behavior; that is, behavior that is inappropriate and ineffective in achieving its purpose.

Adjustment

The term "adjustment" is most frequently used to describe how well a person gets along *in a situation*. The employer will point out well-adjusted employees, the teacher well-adjusted children, in terms of their conformity to the environmental demands of the work or school situation. To a psychologist, however, adjustment implies not mere conformity but a harmonious relationship between the individual and his present environment. A person can achieve adjustment either by adapting his behavior to the requirements of a situation or by changing the situation to meet his personality needs.

Students may show different degrees of effectiveness in adjusting to different areas of living. A student's adjustment cannot be judged entirely on the basis of how he acts in school. For some students, school is the

only area of serious maladjustment; for others, school may be the area of best adjustment, bringing them satisfactions that they cannot achieve in other areas of living.

ADAPTIVE BEHAVIOR AND INTERNAL CONFLICT Redl and Wattenberg warn that, since *adjustment* describes a relationship between an individual and his environment, "adjustment is a reasonable criterion of mental health *only if the demands of a situation are reasonable.*"² One would not wish to apply the term "well adjusted" to the child who adjusts to the code of a gang of child thieves or the adult who adjusts to the requirements of the Nazi regime. Children adapt to the constantly nagging parent or teacher by developing psychological deafness; to the tyrannical parent or teacher by various types of aggression or by passive resistance. These reactions indicate fairly normal adaptive devices—and a need for change in the environmental situation. Such *adaptive* behavior may be essential for the person's "survival" in psychological terms.

Redl and Wattenberg stress that adjustment is concerned with much more than the harmony between a child's *surface behavior* and environmental demands.

Are the things he is doing in harmony with his own feelings? If a person is torn by deep, unresolved conflicts, no matter how docile the behavior, he cannot be considered well-adjusted. Under pressure from home, for example, some youngsters will do phenomenal school work, but become sullen and irritable in the process. . . . A child's happiness is an important clue.³

In other words, an adjusted person "is able to work out good relationships in environments that are in harmony with his own values and do not make unreasonable demands upon him."⁴

THREE TYPES OF FAILURE TO ADJUST TO ENVIRONMENTAL DEMANDS Psychologists have frequently studied the problem of personality through observation of *abnormal* behavior. If the concept of adjustment is viewed from the negative point of view, it is seen that poorly adjusted individuals may be grouped into three general categories, described by Cattell as follows:

Psychotics escape right out of the culture pattern into unreality. Neurotics try hard to conform but do so at the cost of ruinous internal mental conflict.

² Fritz Redl and William W. Wattenberg, *Mental Hygiene in Teaching* (New York: Harcourt, Brace & World, Inc., 1959), p. 169.

³ *Ibid.*, p. 170.

⁴ *Ibid.*, p. 169.

Delinquents prefer to have the conflict between themselves and society. Neurotics and delinquents have in common an incapacity to take the cultural pattern. . . . They differ in that the delinquent is generally under-inhibited and the neurotic over-inhibited.⁵

The neurotic youth has tended to inhibit the impulses that arise from his failure to work out a harmonious adjustment with his culture; the delinquent (sometimes as a result of low intelligence, undesirable neighborhood influences, or failure to develop an effective conscience through early identification with a parent or other adult) satisfies his impulses in ways that bring him into conflict with society.

A WORKING CONCEPT OF PERSONAL-SOCIAL ADJUSTMENT On the basis of the review of concepts just presented, the following working concept of *personal-social adjustment* is presented as an orientation or frame of reference for Chapters 8 and 9.

A person may be described as having good personal-social adjustment who

1. Establishes reasonably harmonious relationships with others in different environmental settings (in home, school, and community) without developing persistent internal conflicts that make him unhappy, dissipate his energy through nervous tension, or result in ineffective behavior.
2. Is able to devote most of his energy to the satisfaction of purposes or goals that he accepts as worthy and that are accepted as such by his culture.
3. Shows a degree of control of emotions and impulses that is typical of his age group; retains the spontaneity, creativity, and willingness to experiment and explore that are essential to further growth; is increasingly able to work toward more remote goals and to accept guidance from others without servility, evasion, or resentment.
4. Conforms sufficiently well to the standards or codes of his own age, sex, and culture groups as to allow himself to achieve a sense of belongingness and to be accepted by these groups.

PERSONALITY DESCRIPTION

For some purposes, such as in vocational guidance, we are not merely concerned with identifying poorly adjusted children. We are interested in *describing* the individual's personality.

If we are going to attempt personality description, there are two possible approaches: (1) the psychometric approach, in which we attempt to obtain a numerical estimate of each of several dimensions or traits of

⁵ Raymond B. Cattell, *An Introduction to Personality Study* (London: Hutchinson's University Library, 1950), p. 182.

personality for each individual, and (2) a more impressionistic or clinical approach, in which we use observation, interviews, and other techniques to obtain clues concerning the individual's needs, problems, and conflicts, and integrate all this evidence into a composite, integrated picture of the person as a whole.

In the area of personality study, the educator or psychologist committed to the psychometric approach tends to administer a standard set of questions, called a personality questionnaire or inventory. He prefers to ask everyone the same questions; and he prefers selection-type questions to those allowing free response. He is especially pleased if he can compare a person's responses with those made by persons in a norming sample, who have answered the questionnaire under similar conditions. He wants to interpret the individual scores he obtains in terms of the findings of well-designed validation studies.

Although his methods seem very scientific, the psychometrician tends to select for study only those aspects of personality on which one can develop objectively scored questions. He tends to disregard significant aspects of personality that elude precise definition. His defense would be that what cannot be defined cannot be measured; that a technique or instrument on which results vary with the examiner or test-user does not provide admissible evidence.

A psychologist who prefers the clinical approach to personality study will prefer free-response questionnaires and test situations in which the person has an opportunity to interpret the question or task as he perceives it, and respond in an individualized manner. The clinical approach admittedly results in data that are not comparable from one person to another and that must be subjectively interpreted. However, the clinical approach can lead to usable hypotheses when the leads produced from a variety of sources tend to reinforce each other and when additional data are obtained in areas in which the findings do not converge.

In contrasting the psychometric and clinical approaches to the study of individuals, Cronbach makes a helpful analogy to concepts in "information theory."

He [the information theory specialist] distinguishes two attributes of any communication system: bandwidth and fidelity. . . .

The classical psychometric ideal is the instrument with high fidelity and low bandwidth. . . . A college aptitude test tries to answer just one question with great accuracy. . . .

At the opposite extreme, the interview and the projective technique have almost unlimited bandwidth . . . the interviewer may cover twenty topics in a half-hour, and note an even larger number of traits. . . .

Bandwidth can be greatly increased when it is possible to confirm or reverse judgments at a later time . . . Narrowband instruments are desired to make

final, irreversible decisions about important matters (e.g. scholarship awards). . . . As a first stage, the wideband test scans superficially a range of important variables, pointing out significant possibilities for further study. In this use the wideband procedure is used for *hypothesis formation*, not for final decisions. . . . *The fallibility of wideband procedures does no harm unless the hypotheses and suggestions they offer are regarded as verified conclusions about the individual.*⁶ [Italics added]

General observation of individuals, interviews that range widely over any topics of concern to the student, and projective tests in which persons give individualized responses to ambiguous pictures or to inkblots have wider bandwidth and lower fidelity.

SOURCES OF DATA ABOUT THE PERSONAL-SOCIAL ADJUSTMENT OF INDIVIDUALS

Self-report

In the preceding chapter on interests and attitudes, we found that the most widely used methods involved *self-report* instruments. In both interest inventories and attitude scales, the test-user summarizes and interprets what the individual has voluntarily reported about himself. In studying personal-social adjustment we also tend to rely on self-report. This appears to be an ideal approach because each person knows more about his fears and hopes, his feelings of adequacy or inferiority, than his teacher or his peers could possibly know. The interview, the autobiography, and the personality inventory are attempts to tap this source of information. The interview and autobiography will be discussed in this chapter, while the advantages and limitations of personality inventories are considered in Chapter 9.

Observation of Relevant Behavior

Another approach is to study *behavior* relevant to the individual's personal-social adjustment. A person may say that he is brave but behave in a cowardly manner; he may say that he is at ease in a social group and yet show obvious signs of tension when observed in a social gathering.

If one could obtain a large, representative sampling of relevant behavior, undistorted by the effects of being observed, one would probably have a more valid basis for judging a person's fearfulness or social poise than could ordinarily be obtained from a self-report inventory. However,

⁶ Lee J. Cronbach, *Essentials of Psychological Testing* (New York: Harper & Row, Publishers, Inc., 1960), pp. 602-604.

as we stressed in the discussion of indirect measurement in Chapter 5, many significant behaviors are not readily observed and do not recur frequently. Moreover, the natural or real-life situations in which we observe individual A are not comparable to those in which we observe individual B.

When we attempt to set up special situations to evoke from each person the behavior we wish to observe (such as a standard situation in which each person is purposely frustrated), we gain in comparability of situations; however, the situation may become artificial and the behavior of the person be modified thereby.

Teachers, however, have unusually fine opportunities for observing students in a variety of *natural* situations—in the classroom, on the playground, and in extracurricular activities. Observation in natural situations involves greater bandwidth with corresponding loss of fidelity. However, as already emphasized, if such techniques are used as a basis for *hypotheses* that are subject to revision, they can prove very useful. Suggestions for the use of observation in both natural and specially designed situations will be considered in this chapter.

Projective Techniques

Projective tests provide rich opportunities for observing behavior; they are especially designed to stimulate the person to behave so as to reveal more of his “real self,” of his partially repressed fears and hopes, than he would if he were on guard. The examiner tries to stimulate his imagination and to encourage him to make free, uncensored responses. A child may be asked to act out dramatic scenes of his own choosing with dolls and miniature props; or a person may be asked to tell a story about each of several pictures. These pictures have been selected as having psychological significance but ambiguous content, that is, pictures that can be interpreted in any one of several ways.

Observing and interpreting an individual's behavior in projective test situations (which stimulate his imagination and reduce his censorship of his responses) requires special training, not only in test administration but in the background of psychological theory required for the formulation and testing of hypotheses. Special courses on these procedures are provided for persons preparing to be clinical psychologists or to work in related fields. Projective techniques will be briefly described in Chapter 9.

The Opinions of Others

As a basis for making certain types of judgments, the opinions of others are necessary. In the selection of students or employees, one cannot depend entirely on self-report techniques; ratings by teachers or employers

on relevant characteristics seem to be essential. Teachers usually wish to communicate to students and parents their impressions concerning the student's work habits and attitudes, his sportsmanship on the playground, and other aspects of his personal-social development. Teachers' ratings on these characteristics are ordinarily included on elementary school report cards.

If we wish to study the social acceptance of individual students, sociometric techniques can prove valuable. A summary of these data will reveal which students are unchosen by others; these students might not have been identified by observation, for they may be "hangers-on" or "fringers" in social groups. Sociometric techniques, rating scales, and other means of obtaining the opinions of others are considered in this chapter.

SELF-REPORT TECHNIQUES

Self-report inventories have been deferred to Chapter 9 because they should be used only by persons with special training in their interpretation. Other self-report techniques that can be used effectively by teachers include: (1) interviews and (2) autobiographical materials.

Interviews

Whenever we plan to use interview data as a basis for ranking students or employees, we become concerned about obtaining comparable data. That is, when interviews are used as part of a selection process (for example, with applicants for scholarships, college admission, or employment), it is essential to agree upon the information needed, the characteristics on which ratings are to be made, and probably to compile a list of questions that should be used.

In many situations, however, we are not attempting to rank individuals. We may want information that will aid in diagnosis and in forming hypotheses about desirable next steps in helping the individual. Hence, the focus of the interview will change from student to student. The teacher or counselor may be especially concerned about Mary's difficulties in gaining acceptance in a social group, with Peter's inability to accept criticism, with Roy's tendency to close his mind to new ideas, or with Susan's feelings of conflict between her desire to go far in a chosen career and her desire to be "one of the gang" among her peers.

If the teacher believes that the student's problems originate in his family relationships, he will skillfully lead the conversation to home responsibilities; leisure activities the student enjoys most and his companions in such activities; activities participated in with mother, father, brothers,

and sisters; how the student usually spends his weekends; whether he has a favorite brother or sister; and the like. These questions should stimulate conversation that reveals the student's underlying *feelings* about his family relationships.

The interview is valuable in working with almost any type of problem. The nature of the interview, its length, and the number of interviews necessary must be individualized to fit the student, his maturity level, and his problem.

PREPARING FOR THE INTERVIEW The teacher or counselor should prepare for an interview by reviewing available data and organizing his own knowledge about the student. The cumulative-record folder should be checked, especially for comments by previous teachers, relevant test data, and home information. It is highly important, for example, to know whether there is a stepparent, whether other relatives live in the home, the number and age of siblings, and other facts that will help to direct the teacher's questions and also increase his awareness of topics on which the student might be unduly sensitive.

If the teacher has prepared for the interview by organizing information about the student's behavior and possible environmental pressures, he will probably have made certain hypotheses about the "why" of any disturbing behavior. Such hypotheses can be of great value in giving direction to the interview and helping the teacher to elicit information that supports or negates his tentative diagnosis. If the teacher is to use his time to greatest advantage, he should make use of such hypotheses. However, he must be alert to other causal factors that he may not have considered, and he must avoid the danger of seizing upon a single explanation and marshalling evidence to support it.

ADEQUATE RAPPORT AND COMMUNICATION The importance of achieving rapport with the student cannot be overemphasized. The teacher will already have laid the foundation for such rapport in his day-by-day relationships with the student. In order for the atmosphere to be friendly and relaxed, many elementary school teachers invite children to help them after school with some interesting activity.

The counselor may be meeting the student individually for the first time. However, he will have had opportunities (through group meetings) to communicate to students that he is a person who will listen to their problems and will help them make their own decisions. On the basis of earlier group contacts that have helped him to develop an appropriate image of his role, the counselor can now communicate his interest in the counselee as an individual.

The effectiveness of the teacher's or counselor's communication to the student depends not only on his use of a simple vocabulary but on his using concrete illustrations instead of generalities. Moreover, it is imperative that the teacher or counselor avoid blocking communication by putting the student on the defensive through moralizing or cross-examination, or by seizing the initiative and dominating the interview situation.

Valid and helpful information is more readily obtained by being an attentive and sympathetic listener than by doing most of the talking. As one gains in experience as an interviewer, one learns to avoid the type of question that can be answered by "yes" or "no" and to substitute questions that stimulate the student to describe significant aspects of his environment and the way he *feels* about them.

Interpreting Information Obtained through Interviews

Because of the subjectivity of the interview technique, it is important that information obtained through this means be interpreted in the light of more objective data from other sources. Students sometimes give information that is distorted or actually false. The student who needs help most may be the least willing to confide; moreover, he may have little insight into the reasons for his own behavior.

Although it is essential to distinguish fact from fiction, the teacher must realize that information on the students' feelings and his distorted perceptions is highly important. A student's feeling that his mother is disappointed in him or that his home is old-fashioned is a significant part of the information needed to help him, even though his perceptions do not agree with the facts as seen by a more objective observer.

Although the teacher must avoid cross-examining the student about his family relationships, he can usually obtain highly significant information through the student's informal conversation and through his replies to seemingly casual questions concerning home duties and routines, family recreation, the ages of his brothers and sisters, activities they share together, and the like. Such information should be sought not out of curiosity but as a basis for understanding the adjustment problems of the student. It is generally recognized, for example, that a student's attitude toward adults in authority, toward imposed tasks, and toward school is conditioned to a large extent by his attitude toward parental authority. It is also recognized that a student's reaction to his classmates may represent a displacement of his feelings toward his brothers and sisters.

At the risk of oversimplification, an attempt has been made in Table 8.1 to summarize the "do's and don't's" of interviews concerned with students' adjustment problems.

Table 8.1**Pointers on Interviews with Students Concerning Adjustment Problems**

DON'T	DO
<ol style="list-style-type: none"> 1. Confuse the real purpose of the interview with the immediate problem that precipitated it. 2. Talk too much or dominate the interview situation. 3. Cross-examine the student. 4. Seem rushed or preoccupied. 5. Moralize or pass judgment on a student's behavior in such a way as to lower his self-respect. 6. Antagonize the student or put him on the defensive. 7. Prod the student into revealing confidential information about his friends or his family that he may later regret having told you. 8. Create a feeling of dependence on the part of the student. 9. Try to accomplish too much in one interview. 10. Judge the value of the interview entirely on the basis of results accomplished at that time. Even though little specific progress seems to have been made, a good interview may have laid the basis for closer teacher-student relationships and later confidences. 	<ol style="list-style-type: none"> 1. Prepare for the interview by obtaining and organizing pertinent information. 2. Help the student to relax through an informal greeting, participation in some after-school activity, and/or discussion of pleasant topics of especial interest to him. 3. Put yourself in the student's place; try to see things through his eyes. 4. Be an attentive listener, watching for leads suggested by the student's conversation. 5. Talk on the student's level, maintaining an attitude of cooperation rather than a display of authority. 6. Shift as much responsibility to the student as he is able to handle, helping him to think through "next steps." 7. End the interview with a forward-looking attitude, leaving the student with less anxiety and greater self-confidence. 8. Make special note of the student's last comments as he leaves the interview situation; he may touch on problems that he wished to discuss but could not find the courage to propose earlier.

After the interview, the teacher or counselor should record the salient points on a simple "Record of Interview" form for filing in the student's folder.

Autobiographical Material

An autobiography may be brief and stereotyped or a highly revealing and helpful document. In order to avoid a routine chronology or listing of facts, the teacher should discuss the writing of a good autobiography with the class. He should read excerpts from autobiographies that involve the writer's problems, worries, feelings, and attitudes. Then he should suggest

the kinds of subjects that students may wish to include in *their* autobiographies, such as places in which they have lived, their parents' wishes and concerns for them, their closest friends, their changing vocational plans, their worries, and other significant topics. Students should be assured that their autobiographies will be read only by the teacher and discussed with no one. If autobiographies are to be filed in the students' folders, students should know that they will be used only by their own counselors.

The writer of an autobiography may choose to omit, to overemphasize, or to distort any aspect of his life. This very freedom of expression, however, makes the report a revealing one to the person who has the necessary training and experience for its interpretation. An incident reported in unusual detail is usually of great significance to the writer. Omission of a period in childhood may indicate merely that things went well during that period; however, the omission of reference to one parent or one sibling *may* be indicative of a strained relationship.

Some autobiographies—in fact, most autobiographies of children below age 14—are a compilation of commonplace facts and of incidents that appear to have no special significance.⁷ However, when an older adolescent writes an autobiography of this type, it may be an indication that he is (1) a highly defensive individual who hides his problems, (2) a person with shallowness of feeling, or (3) a student who was not "sold" on the significance of the assignment or who distrusted the use that would be made of the material.

In reading an autobiography, a teacher or guidance worker should be sensitive to variations in tone or mood. In speaking, a person reveals through his voice, gestures, and facial expression the extent of his concern with his subject; similarly, he may show in his writing (for example, through the use of emotionally charged words) that certain experiences have touched him deeply.

A teacher or counselor who has considerable factual data about a student may be able, in reading his autobiography, to detect misrepresentations of facts or evidence of self-deception. A student, for example, may have avoided facing, or may have distorted, facts about his low achievement, lack of social acceptance, or failure in job situations. Such distortions usually indicate areas in which the student feels especially vulnerable.

Since adequate interpretation of autobiographical material involves great sensitivity and skill on the part of the interpreter, the teacher or counselor working with such materials should increase his insights into this and

⁷ Gordon W. Allport, *The Use of Personal Documents in Psychological Science* (Washington, D.C.: Social Science Research Council, 1942), p. 80.

other child-study techniques through participation in case conferences and other in-service education experiences, and through study of references that discuss the autobiography in greater detail than is possible in general textbooks on measurement and evaluation.

OBSERVATION OF BEHAVIOR

Self-report techniques can, under ideal circumstances, provide exceedingly helpful information obtainable in no other way. It would be unwise, however, to limit our study of an individual to the information he voluntarily reports. We should capitalize on our opportunities to observe his actual behavior in a variety of situations. The observation of actual behavior, and techniques of improving its validity and reliability, will be considered in this chapter section.

Systematic Observation of Behavior

In research studies, systematic observations are often made of a single type of behavior as it occurs in natural situations. A time-sampling plan is typically developed in which each subject is observed in a random sampling of situations. For example, in studying aggressive behavior in nursery school children, we would have to define carefully which behaviors were to be classified as "aggressive." We would have to decide whether "hitting a child and taking his toy" was one or two incidents of aggressive behavior. We might plan to observe each child for several five-minute periods randomly scattered throughout a nursery school session. We would train observers and check on the degree of agreement between observers in their counting of aggressive behaviors (for the same children during the same time interval). In other words, we would do everything we could to increase objectivity and avoid bias, with respect to times observed, interpretation of terms, and many other factors. If we thought that the presence of observers would modify the children's typical behavior, we might observe the children through a one-way-vision screen.

Informal Observation of Behavior in Natural Situations

The teacher is not willing to restrict his range of observation as narrowly as is the researcher. The teacher is interested in any behavior that helps him to understand the child better and give him leads as to how the child can be motivated and guided.

SITUATIONS AFFORDING OPPORTUNITIES FOR OBSERVATION Teachers are able to observe children in a wide variety of situations, to observe the

roles a child plays in different social groups, and to note variations in his behavior from one situation to another. Instead of formally setting aside observation periods, the teacher notes significant incidents whenever they occur. These incidents may occur in various kinds of classroom activities, in the cafeteria, on the playground, and in extracurricular activities. Certain activities, however, are *especially* likely to provide clues with respect to the student's personal-social adjustment.

1. The informal discussion period (sometimes called the "show and tell" period) that characterizes many elementary school classrooms is an excellent time for observing child behavior. Note a child's willingness or reluctance to volunteer, his tendencies to exaggerate or even to fictionize, his need to compete with other children in telling about "bigger and better" exploits. During these discussions, children often reveal valuable information about relationships with parents or siblings, typical leisure activities, home responsibilities, and the like.
2. When students are working on arithmetic practice materials or other "self-directed activities," the teacher can observe their attitudes toward an imposed task, their persistence or distractibility, their dependence on other children or adults, their confidence in their own judgment in activities in which there is no set procedure (for example, word problems or laboratory exercises), and the like.
3. When students are working with others on a group activity, the teacher has opportunities to observe the degree to which each student seems able to show cooperative behavior, the extent to which he dominates the group and the techniques he uses in doing so, his reactions when his suggestions to the group are rejected, and the like.⁸ When students are working on committee activities under the leadership of a classmate, the teacher can observe which students seem unable to take the initiative, follow through on assigned responsibilities, or show good work attitudes unless they are working under the close supervision of adults.
4. Any period in which students discuss controversial issues is an excellent time for observing student behavior. A student's willingness or reluctance to take a stand, his tendency to defend his own ideas just because they are his, or to accept or reject ideas because he likes or dislikes the individuals proposing them, his need to compete with other students and to win in any discussion—all are significant clues to his adjustment.
5. When a teacher sponsors a club or other extracurricular activity, he has the opportunity to observe students in their spontaneous social groupings and to obtain more valid data concerning those who are popular or socially isolated than he can obtain in the classroom. Students who seem at ease in the relatively formal atmosphere of the classroom may show shy, withdrawing behavior or aggressive, exhibitionistic behavior when participating in informal social activities with their peers.
6. Students' reactions in role playing and other types of dramatization are often revealing. The characters they choose to play, or those assigned to them by their peers, may be significant. As students participate in such

⁸ Gertrude Driscoll, *How to Study the Behavior of Children* (New York: Bureau of Publications, Teachers College, Columbia University, 1941), p. 15.

- impromptu dramatizations, the teacher can note which students take the initiative, which seem to be too tense and inhibited to participate in role playing, and which are characteristically assigned to background roles. Children whose tenseness makes them withdraw from creative self-expression may be able to participate in dramatizations using hand puppets or shadow play.⁹
7. Observing a student's behavior in creative art activities may help the teacher to achieve greater understanding of him. Using art materials that demand less skill opens the channels of creative expression to a larger number of students than would otherwise be possible. Finger painting is considered especially valuable as a technique requiring little skill, permitting great flexibility, and stimulating freedom of expression. Observation of children at work and their *voluntary* interpretations of their art productions may contribute to greater teacher understanding of the child.
 8. Observation of children's behavior on the elementary school playground is an essential part of any study of their personal-social adjustment. When playground activities are not directed by adults (as during the lunch period or the period before school opens), highly competitive social situations may exist in which children show primitive types of aggression, some children are unmercifully teased or are denied use of equipment, and timid children seek refuge from the overwhelming energy and aggressiveness of their more active classmates. In *directed* play situations at the elementary level, and in physical education activities at the high school level, the teacher has the opportunity to note energy level, physical coordination, proficiency in the physical skills so important for social acceptance, self-assurance, and leadership ability.

LEARNING TO DESCRIBE, RATHER THAN JUDGE Making a judgment about a pupil's laziness, cooperativeness, or some other characteristic is not justified unless the observer has followed the procedures used in research studies in defining the characteristic being studied and in obtaining large, representative samplings of behavior. As a rule, however, teachers do not need to appraise each student with respect to his status on certain personality dimensions. They are more interested in formulating better hypotheses about why a child behaves as he does.

In the formulation of such hypotheses, *descriptions* of the ways in which a child reacts to specific situations are most helpful. Instead of labeling a student as "cooperative" or "uncooperative," "interested" or "uninterested," the teacher should note the situations in which the student was cooperative or uncooperative or the activities in which he showed the greatest or the least interest.

RECORDING OBSERVATIONAL DATA Casual, unrecorded observations are interesting but may be misleading as a basis for diagnosing a student's needs. More helpful diagnostic leads will be obtained through the use of

⁹ A. G. Woltmann, "The Use of Puppets in Understanding Children," *Mental Hygiene*, vol. 24 (July 1940), pp. 445-458.

a behavior journal, in which repeated observations are recorded so as to provide a sampling of the student's behavior on his good days as well as his bad days. These observations should be made during class discussions as well as during supervised study; in student-directed as well as teacher-directed activity. Insofar as the teacher's opportunities permit, they should be made in the cafeteria, on excursions, and in extracurricular activities, as well as in the classroom.

The most valuable observational records are descriptions of significant incidents in the life of the student. These descriptions are frequently termed *anecdotal records*. In the anecdotal record, the teacher describes an incident, setting forth briefly and objectively the actual happenings, the setting of the incident, and (if desired) his own interpretation of the significance of the behavior. The teacher's interpretation should be separate from the description of the incident and should always indicate whether or not the behavior is typical of the student. The date and time of day should always be indicated, as well as the setting of the incident and the type of group activity.

The following anecdotal record for a seventh-grade boy is cited as illustrative of the characteristics of such records.

February 25

SETTING During the noon period when the weather is cold, the pupils usually spend their time in the gymnasium playing games.

THE INCIDENT David frequently remains at school although he lives only a short distance from school. Today, as usual, he entered none of the games. He stayed close to the stage, jumped off it several times and rolled around on it. (The stage is at one end of the gym.) He made no effort to join in any of the games. When asked why he did not play with some of the other boys, he replied, "They don't want me."

INTERPRETATION Because of his small stature and his physical condition, David cannot compete with the boys of his own age. The group to which he would like to belong has not accepted him.¹⁰

The teacher who is beginning to write anecdotal records will find the following suggestions practical:

1. Start by selecting one or two students for intensive study.
2. Describe as many significant incidents each week as possible.

¹⁰ Theodore L. Torgerson, *Studying Children* (New York: Holt, Rinehart and Winston, Inc., 1947), pp. 88-89.

3. Do not try to interpret every incident. Make a summary analysis at convenient periods and look for developmental trends in behavior.
4. Concentrate on describing those types of behavior which you believe to have a bearing on the student's difficulties.

As a teacher becomes sensitive to the many symptoms students show in their behavior every hour of the school day, his problem becomes one of selecting those that are the most significant and that justify the time spent in recording. The focus of the teacher's observation will vary from student to student according to the individual's major problems and the teacher's hypotheses concerning possible reasons for his difficulties.

Teachers who systematically record their observations of student behavior develop invaluable records that not only help them in understanding individuals but serve as a basis for conferences with parents and with professional staff members as well.

Observation of Behavior in Specially Devised Situations

When we observe student performance in the skills as a basis for evaluation, we attempt to set up standard situations so that students will be performing under comparable conditions. In basketball, for example, we specify where the student shall be standing when he shoots baskets; in hurdle racing, we specify how high the hurdles shall be and how far apart. When personnel officers in industry and the armed services attempted to appraise different aspects of personal-social adjustment, they also tried to standardize stimulus situations.

SITUATIONAL TESTS One type of specially devised situation that has proved especially promising is the leaderless group discussion. A small group of applicants for teaching positions, for example, might be assigned an interesting and fairly controversial topic, such as the advantages and limitations using television or team teaching. No one is assigned responsibility as moderator. During the course of the discussion, data are recorded concerning the number and nature of each individual's contributions, the leadership pattern that develops, each person's reactions to disagreement with his point of view, and the like. It is obvious that this technique can provide as much information on a number of variables as could be obtained in informal observations of each person over a comparatively long period of time.

Various other "situational tests" involving teamwork in the solving of a difficult problem, reactions to continued frustration and criticism from co-workers, and the like have been used in research studies, and as one basis for selection and classification of employees. Essentially, a situational test

places the subject (or subjects) in a situation simulating the real-life situation in which we would like to observe his behavior. It is similar in many respects to the work sample test, discussed in Chapters 5 and 12; however, the criterion behavior we are attempting to sample is even more complex and more difficult to interpret than that studied in evaluating performance in the skills.

SPECIALLY DEvised SITUATIONS IN THE CLASSROOM The use of situational tests as a basis for *ranking* individuals on personality dimensions requires careful planning and is very expensive. Teachers, however, can utilize specially devised situations in eliciting behavior of a type they especially wish to observe. For example, a teacher can assign students to work together on some fairly difficult project; he can assign individuals to various responsibilities on a classroom newspaper or other classroom project. Or he can use an unfinished film or "reaction story" to stimulate students to role play alternative endings to the story.

If reaction stories are used, they should be carefully selected to meet several criteria. A reaction story should lend itself to good oral reading, being sufficiently dramatic to hold the interest of the students and yet realistic enough to be plausible; it should involve characters with whom students can easily identify themselves; and it should present a genuine problem or issue, to which students of this age would offer and defend a variety of solutions.¹¹ The following résumé of a reaction story indicates that it would meet these criteria:

THE BLIND FISH

Developmental task: *respect for authority*. Children growing up in a society have to learn to curb many of their personal desires and impulses because of the regulations necessary to successful group life.

The basic situation of this story is that of a boy who does not obey rules set down for the welfare of his whole group. A dozen boys at camp go out on a hike with a counselor. They explore a big cave. The counselor lays down some rules for their safety: they must stick together and follow him closely, for if anyone gets separated from the group in this underground maze, he might be lost for days. There is danger, too, of falling off ledges and sliding into pits. The hikers pass an underground pool. One boy induces another to lag behind with him. They wade into the water and their lights disclose fish in the clear pond. They catch some and discover that the fish are blind, and grow excited over their unique find. Then one boy steps into a deep hole and goes under

¹¹ George Shaftel and Fannie R. Shaftel, *Role Playing the Problem Story* (New York: National Conference of Christians and Jews, 1952).

water. He cannot swim and would have drowned if the other boy had not been along to save him. Even the good swimmer, however, gets a cramp in the cold water and barely manages to reach the bank. The crowd returns. Hearing about the blind fish, they all want to catch some. The counselor refuses to permit it. They clamor that he's not fair; the first two boys had caught some of the fish. So why can't they all?¹²

The teacher may prefer to have individual students write out their story endings and hand them in. This procedure has the advantage of obtaining reactions from all students; moreover, the reactions of individuals are not influenced by the positions taken by class leaders. Written statements that appear to be especially revealing can be filed for further reference. Written answers, however, may involve less spontaneity; and for students who do not write fluently, the reactions may be brief and conventional.¹³

OBTAINING THE OPINIONS OF OTHERS: TEACHER RATING SCALES

We have already considered two major sources of information about the individual's personal-social adjustment: (1) self-report through interviews and biographical materials and (2) observation of behavior (in natural and in specially devised situations). The methods we will consider in the remainder of this chapter are based upon the opinions of others. These opinions are most valid when they are based on adequate observation. That is, teachers' ratings and peer ratings are valid and reliable only to the extent that they are based on large and representative samplings of observations. Ratings made by teachers and peers on students they do not know well or on characteristics that are not evidenced in observable behavior tend to have low reliability and validity.

¹² *Ibid.*, pp. 41-42.

¹³ Although they may lack some of the stimulus value of the reaction story, certain story titles may be assigned that will stimulate students to indulge in fantasy. For example, the teacher might assign the title "If I . . .," which the student may complete in any way he wishes, for example, "If I Were a King," or "If I Had a Million Dollars." Other stimulating titles might be: "When I'm Through School," "My Day-dreams," "If I Could Have Three Wishes," "If I Could Have My Way," and the like. Further suggestions concerning theme or story topics that stimulate self-expression, as well as the use of open-ended questions and incomplete sentences, are given in many textbooks on guidance, as well as in Hilda Taba and others, *Diagnosing Human Relations Needs: Studies in Intergroup Relationships* (Washington, D.C.: American Council on Education, 1951).

Observation as a Basis for Evaluative Judgments or Ratings

Teachers are often asked to record judgments concerning student behavior on a rating scale of behavior traits. In fact, the report card is actually a rating scale in which numbers or letters are used as a method of communicating teachers' judgments to students and parents. Obviously, ratings on responsibility, sportsmanship, and other characteristics are of little value unless they are based on careful, unbiased observation of student behavior.

If classes are too large and daily schedules too crowded for the teacher to cumulate a large number of anecdotal records for each student, a realistic compromise is presented by such a record sheet as that developed for the *Personal and Social Development Program*.¹⁴ This record form is organized to focus teacher attention on behavioral evidence of four "personal traits" and four "social traits" of significance in pupils' personal-social adjustment at school. They are as follows:

PERSONAL TRAITS	SOCIAL TRAITS
1. Personal adjustment	5. Social adjustment
2. Responsibility and effort	6. Sensitivity to others
3. Creativity and initiative	7. Group orientation
4. Integrity	8. Adaptability to rules and conventions

In order to help teachers to define each of the "traits," several positive and negative examples of behaviors classifiable under that trait are included. When a teacher makes a brief dated entry of a "critical incident," he notes the code number of the item. Examples for the trait "sensitivity to others" are cited as illustrative:

Sensitivity to Others**BEHAVIORS TO BE ENCOURAGED**

1. Saw that others were not left out
2. Cheered up, complimented, or encouraged others
3. Was kind to someone with handicap or special problem
4. Tactfully provided something for needy child
5. Did something for person not feeling well
6. Corrected or made suggestions to another in tactful manner
7. Interceded for or stuck up for another

BEHAVIORS NEEDING IMPROVEMENT

1. Left another child out of activity
2. Referred to another's race, religion, or nationality in a disparaging manner

¹⁴ John C. Flanagan, *Personal and Social Development Program* (Chicago: Science Research Associates, Inc., 1956).

3. Called another child names
4. Made fun of or teased another about handicap
5. Laughed at the mistakes of others
6. Used sarcasm and disparaging remarks in making criticisms and suggestions to or about others

Ideally, a teacher should record *many descriptive nonjudgmental observations* in a behavior journal before making *judgments* about student behavior. Then when *judgmental ratings* are needed as a means of shorthand communication to students and parents, he can review his cumulated anecdotal records and arrive at summary judgments.

Designing Rating Scales

The first step in the development of a rating scale is the selection and definition of the traits to be rated. The traits listed on a report card or rating scale should be:

1. Independent, in the sense that overlapping should be avoided (for example, responsibility and reliability should not both be listed, for they are too closely related);
2. Definable in terms of observable student behavior (as in the examples given above from the *Personal and Social Development Program*);
3. Related to the major goals of the school program;
4. Reasonably homogeneous or unitary, that is, the trait does not involve component behaviors that have low intercorrelations.

An example of a trait that violates the last criterion is one included in a school-district rating scale under the heading "He does his written assignments." The trait description for the highest level reads: "Always completes written assignments on time; has them well organized and in good form." The three components included (punctuality, organization, and form) would not appear to be highly correlated with each other; hence difficulties in the rating process and ambiguity in communication result.

On most rating scales, the rater's judgment of the student is indicated in one of three ways: (1) by a *numerical rating*, (2) by a check at any point on a line that best describes the degree of the trait possessed by the student (*graphic rating scale*), or (3) by a *descriptive term* that most closely describes him.

Scales requiring a simple numerical rating are used in Chapter 12. Graphic rating scales permit the assignment of any intermediate rating, for example:

Low

High

√

Neither the numerical nor the graphic type, however, help to *define degrees of the trait* being measured.

1. From the American Council on Education rating scale for prospective college students

Does he get others to do what he wishes?

Probably unable to lead his fellows	Lets others take lead	Sometimes leads in minor affairs	Sometimes leads in important affairs	Displays marked ability to lead his fellows; makes things go
----------------------------------------------	--------------------------	-------------------------------------------	-----------------------------------------------	--------------------------------------------------------------------------------

2. From a "Rating Scale for Evaluating Work Habits and Skills" (Long Beach, California, City Schools)

	A	B	C	D	F
He follows instructions	Attends to written and oral in- structions; follows directions accurately		Sometimes lets his attention wander; usually follows directions		Pays little attention to in- structions; follows directions reluctantly or not at all

3. From the Haggerty-Olson-Wickman Behavior Rating Schedules

How does he accept authority?

Respectful, complies by habit	Entirely resigned, accepts all authority	Ordinarily obedient	Critical of authority	Defiant
-------------------------------------	---------------------------------------------------	------------------------	--------------------------	---------

4. From a rating scale for nurses used by the University of Michigan School for Nursing

ADJUSTMENT TO SITUATIONS	Some- times at a loss in familiar situa- tions	Slow to adapt to new sit- uations	Learns new ar- range- ments fairly soon	Quick to adjust to new routine	Very quick to respond to emer- gencies
--------------------------------	---------------------------------------------------------------	--------------------------------------------	--------------------------------------------------------	-----------------------------------------	----------------------------------------------------

Fig. 8.1 Illustrative Items from a Variety of Rating Scales

Descriptive rating scales are more difficult to construct but have greater communication value. If the trait descriptions are carefully worded so that the terms on the negative end of the scale do not sound too damaging, they may also help to distribute ratings more widely over the scale. In Figure 8.1 are presented illustrative items from several descriptive rating scales. In each of these the traits are reasonably homogeneous and are described in behavioral terms.

Since it is difficult for a teacher to make evaluative judgments on all traits for all students, it is advisable for the rating scale to include such a category as "instructor uncertain," or "no opportunity to observe." Inclusion of such an option helps to increase the accuracy of teacher ratings. Space for comments on illustrative behavior or for other supporting evidence is another desirable feature of some rating scales.

Improving the Rating Process

In rating, as in all other evaluation techniques, the teacher is concerned not only with the quality of the instrument used but with the validity and reliability of the actual scores or ratings obtained. Hence, the *process* of assigning ratings is important. In using any rating scale (for example, the "personality and character" section of the report card), the teacher should rate *all* students on a single trait, preferably by sorting the cards into groups representing "high," "average," and "low" or whatever rating categories are being used. After the sorting has been completed, the teacher can reexamine his judgments concerning the students assigned to each category to see whether any students should be reassigned to higher or lower ratings. Ratings are then entered for the trait, and the same procedure is repeated for each trait of the rating scale.

This process is suggested to minimize "halo effect"—that is, the effect of the teacher's general or over-all impression of a student on his rating of the specific traits. If a student rates unusually high or low on personality traits that are important to the teacher, the teacher's ratings on his other traits are likely to be affected positively or negatively by this "general impression." The sorting procedure just described will help to minimize this source of error. Also, as the teacher sorts cards, he will note the relative number of students he has assigned to each category and remember that approximately as many students should be assigned below-average as above-average ratings. Thus he can minimize the generosity error, or the bias toward high ratings evident in the results of most rating procedures.¹⁵

¹⁵ In the absence of knowledge to the contrary, it is best to assume that students are normally distributed with respect to a given trait. Hence, if there are three steps

OBTAINING THE OPINIONS OF OTHERS: SOCIOMETRIC TECHNIQUES

We have studied techniques concerned with how the student sees himself, how he behaves in natural and specially devised situations, and how his teachers record summary judgments on rating scales. The techniques in this final section are concerned with how the student's classmates see him and with his degree of social acceptance by them.

Although the teacher may gain considerable information about students' social relationships through observation, results of class elections, and the like, only sociometric techniques reveal how students would *like to* associate and how their wishes compare with the attitudes of other students toward them.

Selecting the Questions

The first step in making a sociometric study is to choose questions that will stimulate students to reveal their true feelings about other members of the class. Questions of the following types are appropriate:

1. Whom do you wish to sit next to in the classroom?
2. With whom would you like to work on a committee?
3. Whom would you like as companions on a class project?
4. Who are your best friends?
5. With whom do you like to associate after school?

Of the five questions listed above, the first three are ones in which the choices can actually be put into effect. Students are likely to respond more frankly if they feel that they will benefit from an honest report of their feelings. Questions 4 and 5 provide information about close interpersonal relationships but indicate no reason for seeking the information.

The questions also differ with respect to the basis for choice. The first and fourth questions imply friendship and pleasure in proximity. The second and third may bring in the additional elements of interest, skill, work habits, and the like. The fifth question, although it involves personal friend-

on a rating scale, approximately 20 percent should receive the highest rating, 60 percent the middle rating and 20 percent the low rating. If the rating scale is a four-step one, the percentages should be approximately 11 percent each in the lowest and highest categories, and 39 percent in each of the middle categories. With a five-point scale, the percentages are those so often cited for grading "on the curve," that is, 7 percent, 24 percent, 38 percent, 24 percent, and 7 percent. If six or more steps are used, the percentages corresponding to six or more equal divisions of the base line of the normal curve are obtained in a similar way.

ship, may be conditioned by such factors as living in the same neighborhood, membership in clubs, and skill in athletics.

All the questions or criteria listed above involve *positive* reactions, or attractions. Negative questions implying rejection have occasionally been used. Although their use in a research study can be justified, their use by teachers probably cannot. An observant teacher is usually aware of students who are actively rejected and need not focus class attention on them by asking for negative choices. Requesting, or even seeming to condone, negative choices seems contradictory to the teacher's positive expectation that each student should be willing to work with, or sit by, any of his classmates.

If one wishes to obtain reliable data concerning the relative social acceptance of individuals, it is desirable to request five choices rather than a smaller number.¹⁶ From the third grade on, children seem to experience little difficulty in making five choices. Children in grades one and two are usually able to make at least three choices for each criterion.¹⁷

Administering the Questions

The sociometric test should be administered in an informal and natural manner. Ordinarily, the teacher distributes 4 in. by 6 in. slips of paper on which the student writes his name and lists his numbered choices. If more than one question is used, duplicated forms may be advisable. Although the teacher will wish to give the directions informally rather than read them, he should think through his presentation carefully and know what he wishes to say. Jennings has summarized the essential pointers for good administration as follows:

Teachers should always feel free to answer any questions that may occur to the group, both before and during the writing, and should treat the occasion in a business-like manner. The most important things to remember about administering the test are: (1) to include the motivating elements in the introductory remarks, (2) to word the question so that children understand how the results are to be used, (3) to allow enough time, (4) to emphasize *any* boy or girl, so as to approve in advance any directions the choice may take, (5) to present the test situation with interest and some enthusiasm, (6) to say how soon the arrangements based on the test can be made, and (7) to keep the whole procedure as casual as possible.¹⁸

¹⁶ Use of five choices has been found to result in the most stable sociometric data according to research studies summarized in Norman E. Gronlund, *Sociometry in the Classroom* (New York: Harper & Row, Publishers, Inc., 1959), p. 48.

¹⁷ *Ibid.*, p. 48.

¹⁸ Helen Hall Jennings, *Sociometry in Group Relations* (Washington, D.C.: American Council on Education, 1948), p. 16.

Teachers may find that providing a duplicated list of classmates' names lessens confusion, discourages discussion, and reminds students of absences. If desired, such a list may be numbered, with students recording the numbers, rather than the names, of classmates chosen.

Constructing the Sociogram

The general procedure in drawing a sociogram is to locate individuals on a chart so that the "stars" (the most frequently chosen) are near the center; the "isolates," or unchosen, are on the periphery; and other students are located so as to minimize the number of long lines and the number of intersecting lines.

As a general principle, it is well to start the sociogram for either girls or boys by drawing in and labeling the symbols for those who are "stars" and for their mutual friends. Figure 8.3 is a sociogram depicting the boys' choices, as summarized in the tabulation form (Figure 8.2). The first symbols entered on this chart would be those for John A. and Harry E. (the most frequently chosen boys) and their mutual friends, Raymond F. and Roger B. Symbols for additional boys chosen by this group are entered next (Fred D., George G., and Walter H.). The symbols are placed so as to minimize long and intersecting lines. Finally, symbols and arrows for all boys still uncharted are entered, the decision on placement in each case being based on choices given and received. Isolates should be located near the periphery of the chart and placed so that lines can be most easily drawn to their choices.

Figure 8.3 shows that John A. and Harry E. are "stars," each receiving six choices. John's choices of Harry and Raymond are mutual. Although Raymond received only four choices (two of them from fringers in the group), he occupies a position of social importance in the class as a mutual friend of John A. and the first choice of Harry E. Andrew I. is the only isolate in the strict sense of the word; however, Peter C. received only one choice (from the isolate Andrew) and had none of his three choices reciprocated. Barry J. received only one choice (a third choice from Walter H.). He made only two choices, failing to use his full quota; neither of the choices he made was reciprocated.

The boys of this class look for leadership to two boys who are mutual friends, John A. and Harry E. These two boys include as a close personal friend Raymond F., whose only other choices, however, come from two of the less popular boys. All the boys in the class are tied, although somewhat loosely, into one psychological network. However, there are few mutual choices; the boys seem to be held together as a social group chiefly through the mutual friendship of the two "stars."

Chosen Chooser	John A	Roger B	Peter C	Fred D	Harry E	Raymond F	George G	Walter H	Andrew I	Barry J	Intersex Choices
John A		2			1	3					
Roger B				1			2				3- Mary
Peter C				3	1	2					
Fred D	2				3			1			
Harry E	3	2				1					
Raymond F	2						1	3			
George G	1				2			3			
Walter H	2				1					3	
Andrew I	1		3			2					
Barry J*					2		1				
Chosen as 1st choice	2	0	0	1	3	1	2	1	0	0	0
2nd choice	3	2	0	0	2	2	1	0	0	0	0
3rd choice	1	0	1	1	1	1	0	2	0	1	1
Total	6	2	1	2	6	4	3	3	0	1	1

*Did not make a third choice.

Fig. 8.2 Sociometric Tabulation Showing the Choices of Ten Boys.

Interpreting Sociometric Data for Individuals

In interpreting sociometric data for individual students, the teacher is naturally most concerned about the isolates (those students who are not chosen) and the neglectees (those who are very infrequently chosen even when one requests five choices on two or more criteria).¹⁰ Several questions about the isolates and neglectees need to be asked before the seriousness of their isolation and the suitable remedial procedures can be determined.

¹⁰ If five choices are requested on a single question (or criterion) a pupil with one choice is considered a "neglectee" because there are only two chances in 100 that he would receive so few choices if only chance were operating. Similarly, a pupil with nine or more choices would be called a "star." If five choices on two criteria are used, a neglectee is one receiving four or fewer choices; a star, one receiving sixteen or more. If five choices on three criteria are used, a neglectee is one receiving nine or fewer choices, while a "star" is one who receives twenty-two or more. Urie, Bronfenbrenner, and Henry Ricciuti, "The Appraisal of Personality Characteristics in Children," in Paul H. Mussen, ed., *Handbook of Research Methods in Child Development* (New York: John Wiley and Sons, Inc., 1960), pp. 770-817.

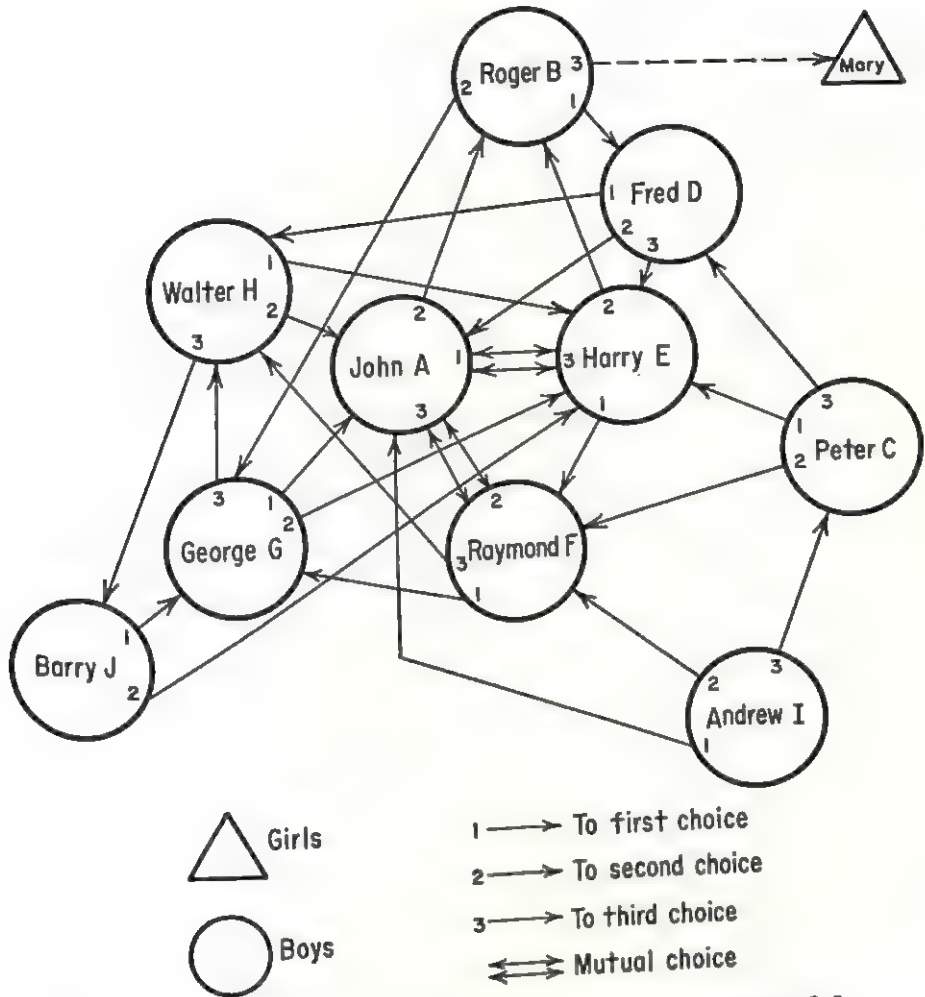


Fig. 8.3 A Sociogram Based on the Data Shown in Figure 8.2

1. Is the student who is isolated in class also isolated in home and neighborhood groups, or is his position in one social group counterbalanced, in part, by his position in another? (This phenomenon frequently occurs when the basis for isolation is ethnic group or social class.)
2. Is the student relatively new to the class?
3. How old is the student? (If a student in the fifth grade or above is isolated, his isolation is probably due to group attitudes toward him, while an isolated child in kindergarten or the primary grades may simply be a shy child who is ignored.)
4. Is the isolate realistic about his position? (Some isolates are quite realistic about their social status in the class, indicating their desire to associate with one or two students with whom there is some basis for establishing friend-

ship. Other isolates are quite unrealistic, listing the most popular members of the group as friends.)

Among the isolates and neglectees, the teacher is almost certain to find students with adjustment problems that he might otherwise have overlooked. The teacher will probably find among the "stars" some happy, likable students who rank high in many aspects of personal-social adjustment. The fact that a student is a "star," however, does not necessarily imply optimal personal-social adjustment. Among individuals who make persistent efforts to achieve popularity and leadership will be found some whose drive arises out of deprivations in other areas of adjustment—for example, affectional relationships in the family. These relationships emphasize again that data from several approaches or techniques must be combined if the teacher is to achieve an understanding of the student and his adjustment problems.

Validity and Reliability of Sociometric Data

If students are asked sociometric questions and promised that their choices will be considered in assignment to seats or to class committees, they are likely to give their actual preferences (as of that time and with respect to the questions used). In a sense, their choices are criterion data; and we need not be concerned with relevance.²⁰ Our only concern appears to be with reliability.

Data concerning the stability of sociometric choices are encouraging. Gronlund²¹ found an average stability coefficient of .75 over a four-month interval when he studied groups of children in grades 4–6; while Bonney,²² who has conducted several studies on stability of choices, found that stability coefficients obtained over a one-year period ranged from .67 to .84. Requesting five choices for two or more questions will yield sociometric scores of considerable stability, if the study is made in the upper elementary

²⁰ If the sociometric questions are ones that do not imply fulfillment of choices, such as "Who are your best friends?" we cannot take validity for granted. In one research study, however, Byrd found correlations of approximately .80 between (1) the number of choices each student received when classmates were asked to list those they preferred to work with on a class play and (2) the number of choices each student received when "real life" choices were made.

²¹ Norman E. Gronlund, "The Relative Stability of Classroom Social Status with Unweighted and Weighted Sociometric Status Scores," *Journal of Educational Psychology*, vol. 46 (October 1955), pp. 345–354.

²² M. E. Bonney, "The Constancy of Sociometric Scores and Their Relationship to Teacher Judgments of Social Success and to Personality Self Ratings," *Sociometry*, vol. 6 (November 1943), pp. 409–424.

grades or above. Studies made in the kindergarten and primary grades show somewhat lower reliability.

If "total choices received" is presumed to represent the construct of "general social acceptance," we must be concerned with the amount of consistency in the student's social acceptance from one social group to another, and from one criterion to another. Gronlund and Whitney²³ obtained an average r of .72 between number of choices each student received as seating companion (within the classroom) and the number he received as future classmate (throughout the school).

When the consistency of sociometric status from one criterion to another has been studied, the correlations have naturally varied in some degree with the similarity of the criteria. In a study of 1258 sixth-graders, by Gronlund,²⁴ five choices were requested for each of three criteria: seating companions, work companions, and play companions. The intercorrelations among criteria ranged from .76 to .89. The highest correlations were between number of choices received as seating companion and as work companion; the lowest correlations were between work and play criteria. Northway suggests the following hypothesis:

An individual's acceptance score as measured in one group is a reliable index to what his acceptance score will be in a reasonably similar (culture-age) group. That is, his acceptance score is an outward measure of a psychological characteristic called acceptability.²⁵

This hypothesis should not, however, be interpreted as an assurance that data from any sociometric testing program may be taken as valid evidence of a pupil's general social acceptability, unless the following conditions are present: (1) the pupil has been well motivated in taking the sociometric test; (2) there have been opportunities within group situations for building up social interrelationships; and (3) the criteria or questions used were designed to reflect *general* social acceptance, such as choice of seatmates or friends, rather than choice of pupils for a relationship involving special qualifications, such as pupils "to help you in arithmetic" or choice of pupils for such a position as class secretary.

²³ Norman E. Gronlund, and A. P. Whitney, "Relation between Pupils' Social Acceptability in the Classroom, in the School, and in the Neighborhood," *School Review*, vol. 64 (September 1956), pp. 267-271.

²⁴ Norman E. Gronlund, "Generality of Sociometric Status over Criteria in the Measurement of Social Acceptability," *Elementary School Journal*, vol. 56 (December 1955), pp. 173-176.

²⁵ Mary L. Northway, *Appraisal of the Social Development of Children at a Summer Camp*, Psychology Series, vol. 5, No. 1 (Toronto: University of Toronto Press, 1940).

Interpreting Data on Group Structure

An important function of sociometric techniques is to reveal the extent of integration or cohesiveness of a group. The following questions will assist the teacher in appraising the social structure of his class:

1. What is the leadership structure within the group? If there are two or more well-organized groups, what are the attitudes of their leaders toward each other? Is the division into groups based on cleavages that appear in adult society (ethnic group, nationality, and social class)?
2. How integrated or cohesive is the class social structure? If questions are used that permit outside-class choices, what percentage of choices were made within the class? To what extent have mutual choices been made? Are there any groups that are isolated from, or rejected by, the rest of the class? Are there any evidences of maladjustment or delinquency in these isolated or rejected groups?
3. What do the majority of most-chosen students have in common (high social status, religious denomination, length of residence in the community, participation in after-school activities, and the like)? What do the neglectees have in common (nationality, lower socioeconomic status, newness to community, or residence in trailer camp, housing project, or orphanage)?

In the analysis of a sociogram, the age of the students must always be considered. In the kindergarten and primary grades, one is likely to find a relatively large number of isolates and a low number of mutual pairs. Choices are highly unstable, especially in the kindergarten and first grade.

In working for reorganization of social groups, the teacher may find it especially useful to request data from students on both their interests and their choices of friends. Such a procedure allows the teacher considerable latitude in the formation of committees, for he may safely disregard the first and second choices of associates made by a student leader in the interests of honoring his first choice on activities. In this way, isolates can be placed with students with whom they would like to associate; clique members can be distributed among several committees on the basis of choices they direct toward nonclique members; clique leaders can be placed in a position of responsibility on committees involving several members of another clique; and a variety of patterns may be tried out. The teacher can probably be somewhat more venturesome in arranging short-lived committees for a party than in planning long-term committees involving earnest cooperation and serious work on some classroom project.

In satisfying the friendship choices of individuals, the following principles have been found to be justified in practice:

1. In order to carry out as many expressed wishes as possible, it is generally best to start with the children who have not been chosen at all or only seldom. It is usually better to give an unchosen pupil his own first choice. For example,

if David chooses Patty first, Lee second, and Willard third, and no one chooses him—then David should be placed with Patty.

2. Give any pupil in a pair relation the highest reciprocated choice from his point of view; his first choice if this is returned; his second if this is returned and his first is not, or his third if this is his only reciprocated choice on his list.

3. If a child has received choices only from people other than the ones he chose, then give him his first choice.

4. Make sure that each child has been placed with at least one of his choices.²⁶

Peer-nomination Techniques

Peer judgments are also used in reputation or peer-nomination questionnaires, in which the student is asked to name classmates whom certain "word pictures" in the test seem to describe. The following directions are illustrative:

Here are some little word pictures of children you may know. Read each statement carefully and see if you can guess whom it is about. It might be about yourself. There may be more than one picture for the same person. Think over your classmates and write after each statement the names of any boys or girls who may fit it. If the picture does not seem to fit anyone in your class, put down no name, but go on to the next statement. Work carefully and use your judgment.²⁷

Each word picture should be a brief description rather than a mere trait name. In research studies it has been found advantageous to use pairs of contrasting items, although they should not appear together in the questionnaire. The following pair is quoted from Tryon's study of peer judgments at different age levels:

1. Here is someone who finds it hard to sit still in class; he moves around in his seat or gets up and walks around.

2. Here is someone who can work quietly without moving around in his seat.²⁸

²⁶ Jennings, *Sociometry in Group Relations*, *op. cit.*, p. 73.

²⁷ Stuart Stoke, "The Social Analysis of the Classroom," unpublished mimeographed report, Division on Child Development and Teacher Personnel, Commission for Teacher Education, January 1940.

²⁸ Caroline Tryon, *Evaluations of Adolescent Personality by Adolescents*, Monograph of the Society for Research in Child Development, vol. 4, No. 4 (Washington, D.C.: National Research Council, 1939).

Space is left after each word picture for the student to list the names of classmates who fit the description. For classroom use, the teacher should avoid the use of word pictures that would be viewed negatively by the peer group. In the example above, the negatively worded statement would be quite satisfactory since variations in restlessness are socially acceptable.

One cannot use the number of mentions a student receives for a specific trait as an index of how he ranks on this trait in comparison with other students. A student who is popular will be mentioned very frequently for the desirable attributes; a student who is rejected will be mentioned very frequently for the undesirable attributes. For this reason, one should not study the results with respect to one trait at a time, picking out this boy as most restless, this one as most friendly, and the like. Rather, the entire pattern of mentions for an individual should be studied for significant clues in understanding and helping him. The findings for a given student represent *intraindividual* differences, as perceived by his peers.

Peer-nomination techniques have proved to be one of the most dependable rating techniques. The number of "raters" is very large. Moreover, students are asked to list only those classmates who fit the descriptions; they are not asked to differentiate among those in the average range. Another advantage of peer-nomination techniques is that a student's peers are in a position to observe his behavior in many informal situations in which he is under no pressure to show outward conformity to adult standards. Hence peers obtain evidence on many interpersonal traits that is not available to teacher observers in more formal situations. Finally, peer nominations are significant in that they represent the environment of peer-group opinion in which the child lives.

Gardner and Thompson have developed an adaptation of the peer-nomination scale that permits comparison of data for different groups, as well as comparisons of individuals in different school classes.²⁹ When the teacher summarizes data for his class he can ascertain *for each pupil* (1) the way he regards each of his classmates as a potential satisfier of two important social needs and (2) the way his classmates view him as a satisfier of these needs.

Before the student begins filling in his reactions to his classmates, he establishes a frame of reference for rating by selecting individuals (inside or outside school) who represent for him most, average (or medium), and least need satisfaction, as well as two intermediate points between medium and each of the two extremes. Then each of his classmates is rated as better or less good than one of these five persons in helping him to satisfy

²⁹ Eric F. Gardner and George C. Thompson, *The Syracuse Scale of Social Relations* (New York: Harcourt, Brace & World, Inc., 1958).

the specified need. In the example given in an explanatory article,³⁰ John chose his mother as the one whom he would *most* like to go to when he is troubled with some personal problem. He chose one of his girl classmates for the least rating and an uncle for medium. Another classmate and a neighbor were assigned to the intermediate positions. The needs, in terms of which students rate their classmates, are as follows:

1. A possible source of aid when troubled by a personal problem (included at all three levels)
2. Someone to help him to do something well so people will praise him (elementary level, grades 5-6)
3. Someone to look up to as an ideal (junior high)
4. A person whose company he would enjoy at a party or recreation (senior high)

The first question is used at all levels; the second question varies with the school level.

When the Syracuse scale has been used by research workers, the summary scores have shown reliability coefficients approximating .90. Mid-scores of ratings the student assigns to others, as well as the midscore of ratings he receives, can be interpreted in terms of percentile norms, based on a norming sample of more than a thousand students at each grade level.

Cautions in Using Sociometric Techniques

The primary purpose of using sociometric techniques is to increase the teacher's understanding of the social relationships existing within the class so that he can help students to improve in social acceptance and in desirable social traits. One of the criticisms that is sometimes made of sociometric techniques is that they encourage the students to think critically of one another and so tend to crystallize antagonisms of a minor and temporary nature, widen cleavages between cliques, and intensify or make conspicuous the rejection by classmates of students who are socially isolated. Obviously, no teacher wants results of this type.

A number of suggestions have already been made that are intended to minimize the possibility of damaging after-effects. These include (1) using questions in which the choices made by students are actually put into effect; (2) emphasizing positive preferences rather than rejections or dislikes; (3) keeping the whole procedure as casual as possible; (4) managing efficiently so that questions and discussion of procedure can be kept

³⁰ Eric F. Gardner and George G. Thompson, "Measuring and Interpreting Social Relations," *Test Service Notebook*, No. 22 (New York: Harcourt, Brace & World, Inc., 1959).

to a minimum; and (5) giving careful consideration to the sociometric data when forming committees and other working groups. In addition, the teacher must be careful to avoid, in both his speech and his behavior, conveying to the students the impression that popularity is the most important sign of personal worth. Finally, the teacher should regard sociometry as an exploratory technique in the study of students; when a student is discovered to be a neglectee, much more information needs to be obtained before the teacher has a sound basis for diagnosis.

SUMMARY STATEMENT

Three overlapping concepts of mental health or personality development were examined in this chapter—those of maturity, normality, and adjustment. The concept of “maturity” is a valuable one for teachers inasmuch as they can readily categorize a student’s behavior as characteristic of his own age group or of a younger group. The term “normality” sometimes implies average or typical; at other times it is used in the sense of health vs. illness. The term “adjustment” is most frequently used to describe a person’s relationships to his environment and therefore is a reasonable criterion of mental health only if the demands of the environment are reasonable. It is important also that a person’s adjustment to environmental pressures be made without developing persistent internal conflicts.

Two major approaches to personality description have been used: (1) the psychometric approach, in which one attempts to obtain quantitative estimates of different personality dimensions, and (2) the clinical approach, in which one uses a variety of techniques to obtain clues concerning the individual’s needs and problems. Ideally, we should use the techniques best adapted to a specific situation, keeping in mind that data from any one source should be used only for hypothesis formulation and that conclusions are justified only when data from several independent sources converge to support an hypothesis.

Data concerning the personal-social adjustment of individuals can be obtained from: (1) self-report (through questionnaires, interviews, or autobiographies), (2) observation of relevant behavior, (3) projective techniques, or (4) obtaining the opinions of others (through rating scales, sociometric techniques, and other means). A variety of these techniques were considered in this chapter except that techniques requiring specialized training and supervised experience were deferred to Chapter 9, namely: projective techniques and one type of self-report, that is, personality inventories.

In interviewing students, valuable results can be obtained when good rapport is established, effective communication is maintained, and the teacher has familiarized himself with the background material. In their autobiographies, students may reveal a great deal about themselves, their interests, fears, wishes, self-concepts, and the like.

Direct observation of student behavior can provide highly significant data, especially during self-directed activities, group activities, such as role playing, and extracurricular activities. The skilled observer distinguishes between observation of behavior and its interpretation. Behavior journals and anecdotal

records help to systematize observations and to provide a useful summary of observations over a period of time.

There are several techniques of appraising the student's personal-social adjustment as evaluated by others. Teachers' observations can be summarized concisely on checklists or rating scales. The desirable characteristics of rating scales were presented and illustrated. Suggestions for improving the rating process were also given. Teachers can construct sociometric charts on the basis of students' answers to one or more questions involving a choice of their classmates (as seatmates, fellow-committee-members, or best friends). As finally charted in a class sociogram, sociometric data aid the teacher in appraising the social structure of his class and in identifying students who need help in achieving satisfying peer relationships. Peer-nomination techniques can be used to advantage in certain situations. However, the instruments must be carefully devised, administered, and interpreted if the results are to be of maximum value.

SELECTED REFERENCES

- ADKINS, DOROTHY C., "Principles underlying Observational Techniques of Evaluation," *Educational and Psychological Measurement*, vol. 11 (Spring 1951), pp. 29-51.
- BASS, BERNARD M., "The Leaderless Group Discussion," *Psychological Bulletin*, vol. 51 (September 1954), pp. 465-492.
- BRONFENBRENNER, URIE, AND HENRY RICCIUTI, "The Appraisal of Personality Characteristics in Children," in Paul H. Mussen, ed., *Handbook of Research Methods in Child Development*. New York: John Wiley and Sons, Inc., 1960, pp. 770-817.
- DRISCOLL, GERTRUDE P., *How to Study the Behavior of Children*. New York: Bureau of Publications, Teachers College, Columbia University, 1945.
- FLANAGAN, JOHN C., "The Critical Incident Technique," *Psychological Bulletin*, vol. 51 (July 1954), pp. 327-357.
- , AND OTHERS, "New Tool for Measuring Children's Behavior," *Elementary School Journal*, vol. 59 (December 1958), pp. 163-166.
- GRONLUND, NORMAN E., *Sociometry in the Classroom*. New York: Harper & Row, Publishers, Inc., 1959.
- HEYNS, R. W., AND R. LIPPITT, "Systematic Observational Techniques," in Gardner Lindzey, ed., *Handbook of Social Psychology*. Cambridge, Mass.: Addison-Wesley Publishing Company, Inc., 1954, Chapter 10.
- HORN, ALICE, AND ALFRED S. LEWERENZ, "Measuring the 'Intangibles' in Education," *California Journal of Educational Research*, vol. 1 (September, November 1950), pp. 147-153, 195-206.
- HUTSON, P. W., "Recent Studies in Character-Trait Rating," *Personnel and Guidance Journal*, vol. 38 (January 1960), pp. 364-368.
- JARVIE, L. L., AND M. ELLINGSON, *A Handbook of the Anecdotal Behavior Journal*. Chicago: University of Chicago Press, 1940.
- LANGDON, GRACE, AND IRVIN W. STOUT, *Teacher-Parent Interviews*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1954.
- MAYO, GEORGE D., "Peer Ratings and Halo," *Educational and Psychological Measurement*, vol. 16 (Autumn 1956), pp. 317-323.

- MORENO, J. L., *Who Shall Survive?* 2d ed. New York: Beacon House, Inc., 1953.
- NORTHWAY, MARY L., *A Primer of Sociometry*. Toronto: University of Toronto Press, 1952.
- PEAK, HELEN, "Problems of Objective Observation," in Leon Festinger and Daniel Katz, eds., *Research Methods in the Behavioral Sciences*. New York: Holt, Rinehart and Winston, Inc., 1953, pp. 243-299.
- PROCTOR, C. H., AND C. P. LOOMIS, "Analysis of Sociometric Data," in M. Jahoda, M. Deutsch and S. Cook, eds., *Research Methods in Social Relations, Part II*. New York: Holt, Rinehart and Winston, Inc., 1951, Chapter 17.
- TABA, HILDA, AND OTHERS, *Diagnosing Human Relations Needs*. Washington, D.C.: American Council on Education, 1950.
- TAYLOR, DOROTHEA, "How to Obtain Autobiographies," *Personnel and Guidance Journal*, vol. 36 (February 1958), pp. 426-427.
- THORPE, LOUIS P., AND OTHERS, *Studying Social Relationships in the Classroom*. Chicago: Science Research Associates, 1959.
- WOLFE, DON M., "Fruitful Long Paper: The Autobiography," *The English Journal*, vol. 45 (January 1956), pp. 7-12, 38.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. What important purposes may be served by pupil interviews at the grade level at which you teach?
2. Evaluate the autobiography as a method of child study.
3. Observe children on the playground, and make a record of behaviors indicative of feelings of self-assurance.
4. Why is observation an important and useful method of studying children? In what ways can observational methods be improved?
5. List several important precautions to be adopted by teachers in writing and interpreting anecdotal records.
6. What are the limitations of subjective techniques of evaluation? How may they be overcome?
7. Summarize a published case history and evaluate it.
8. Obtain from the local high school rating scales used in physical education, industrial arts, homemaking, or other classes. Criticize these in light of the principles developed in this chapter.
9. Which of the following traits would probably be most difficult to rate reliably on the basis of repeated observations? (a) promptness, (b) neatness, (c) leadership ability, (d) integrity, (e) interest in a subject.
10. Imagine that you are to present to your faculty coworkers the sociogram for the boys of your class (shown in Fig. 8.3). Outline briefly your explanation of the techniques, together with any implications suggested by the data.
11. Evaluate sociometry as a method of studying children and adolescents. List its advantages and limitations.

Personality Inventories and Projective Tests

Some psychologists tend to consider personality as an indefinable whole—very complex in nature and not susceptible to analysis. Other psychologists consider this point of view mystical, vague, and of little value in practice. They would define personality from a psychometric point of view as a pattern of traits or ways of reacting to environmental stimuli. Fortunately, an increasing number of psychologists are combining these two views. They recognize that measurement of personality can proceed only through attempting the identification of personality components that are definable, relatively independent, and reasonably homogeneous or unitary in nature. They concede, however, the limitations of present attempts to identify and measure personality components. They concede also the critical importance of the individual's *integrating* his various traits or reaction patterns into a smoothly working, effective whole.

We are presenting in this chapter two quite different approaches to the study of personality, that is, structured personality inventories and semi-structured projective tests. They have been grouped together because their use by unqualified persons is hazardous, and their value depends a great deal on the examiner's background in psychology and his ability to synthesize information from many sources as a basis for hypotheses about the individual under study.

In a sense, these two approaches represent the two extremes of the psychometric vs. the clinical approach discussed in the preceding chapter. Persons who have considerable faith in one of these approaches are likely to distrust the other.

The personality inventories discussed in the first chapter section are called structured inventories. A test is said to be structured when it is so designed that all examinees interpret the items in the same way. Although

there is some ambiguity, and consequently some variation in the subject's interpretation of inventory items, the authors of structured inventories try to minimize such ambiguity.

Projective tests involve a minimum of structuring. In projective tests, ambiguous content, which permits a variety of interpretations, is valued. In interpreting ambiguous stimuli, such as ink blots or vague indefinite pictures, the subject is encouraged to react in a highly individualized manner. He may be told, for example, that "any story will do." The clinical psychologist assumes that the subject tends to project into these ambiguous stimuli his own wishes, fears, and repressed conflicts.

The style with which the subject works with paint or clay, or the extent to which he uses form, color, and shading in the interpretation of an ink blot, is thought to reveal clues concerning his characteristic approach to real-life situations. Much more research is needed to validate the hypotheses now utilized in the interpretation of projective tests. Fortunately, well-qualified psychologists check hypotheses developed from projective tests with data from other sources such as interviews and observations.

The person who prefers projective tests likes to observe the reactions of his subjects in situations in which they have opportunities to be individualistic and creative, where their defenses are down and they are caught off guard. Many clinically oriented psychologists contend that only such approaches produce valid data. The psychometrically oriented psychologists, on the other hand, tend to distrust the subjectivity of interpretation of projective tests and to question the hypotheses used as guide lines in such interpretation.

Most of the personality inventories have been developed, standardized, and studied by psychologists with a psychometric orientation. All students taking a personality inventory are presented with a uniform set of questions; the person's responses are objectively scored according to predetermined keys, and the results are interpreted in comparison with norming samples. As we will see in this chapter, however, the *interpretation* of results from such inventories cannot be routine and objective.

While personality inventories are self-report questionnaires, the projective tests involve observation of the individual's performance as he tells a story, draws a picture of his family, or utilizes various dolls and background settings in dramatic play. Qualitative interpretations of the person's style or method of attack may be more important than any quantitative evaluation of the product. Projective tests present unstructured or semi-structured situations that permit a variety of perceptions and interpretations. In fact, the examiner encourages the individual to allow free rein to his responses and reassures him that his own individualized interpretations are desired.

PERSONALITY INVENTORIES

In the early days of personality measurement, psychologists attempted to develop inventories that measured single aspects of personality. Allport was concerned with ascendance-submission; Marston and others with extroversion-introversion. Soon, however, a number of psychologists attempted to meet the need of psychologists, research workers, and others for a test that would describe many aspects of personality, which would provide a *personality profile* comparable to the achievement profile obtained by the use of an achievement-test battery. A number of self-styled personality inventories were developed, each based on the author's own list of personality traits. Eventually, reactions developed against such test-construction practices. As Spencer said, "Personality traits cannot be created by the psychologist."¹

Studies of Components or Factors in Personality

One of the basic research problems in the measurement of personality is to identify relatively independent personality traits. As one might expect, factor analysis has not met with as good results in personality as in ability measurement. Researches to date have shown less agreement.

Cattell, Guilford, and others have applied the techniques of factor analysis in an attempt to find personality traits that are *unitary* or homogeneous—that is, groupings of specific elements of behavior that tend to go together in individuals, which are functionally interrelated. Behavior can be described in very small units, such as "jumping at the barking of a dog," or "biting one's finger nails," as is done in observational records. However, it would make for economy of description and would certainly facilitate personality measurement if it were established, for example, that "nervousness" were a unitary trait of personality.

Cattell found that there were approximately 5500 personality trait names in the English language. A selected list of 171 personality trait names was first developed by examining the original 5500—eliminating the trivial and grouping together those that were obvious synonyms. All later simplification of the list, however, was done by the statistical procedure known as "cluster analysis." As a result of his studies, Cattell identified 35 *surface traits*; that is, traits that were evidenced in surface behavior and that seemed to include (on the basis of statistical analysis) the character-

¹ Douglas Spencer, *The Fulcrum of Conflict, A New Approach to Personality Measurement* (New York: Harcourt, Brace & World, Inc., 1939), p. 20.

istics present in the original list of 171 qualities. Further statistical analysis revealed *source traits* that were more nearly basic and accounted for many of the surface traits. Of the source traits that he has identified, Cattell considered six to be best established.²

SOURCE TRAIT	CHIEF CHARACTERISTICS
A Good-natured, easy-going, ready to cooperate, attentive to people, soft-hearted, kindly, trustful, adaptable, warm-hearted	vs. Spiteful, grasping, critical, obstructive, cool, aloof, hard, suspicious, rigid, cold
B Intelligent	vs. Mentally defective
C Emotionally mature, emotionally stable, calm, phlegmatic, realistic about life, absence of neurotic fatigue, placid	vs. Lacking in frustration-tolerance, changeable, showing general emotionality, evasive, neurotically fatigued, worrying
E Assertive, self-assured, independent-minded, hard, stern, solemn, unconventional, tough, attention-getting	vs. Submissive, dependent, kindly, soft-hearted, expressive, conventional, easily upset, self-sufficient
F Talkative, cheerful, placid, frank, expressive, quick, alert	vs. Silent, introspective, depressed, anxious, uncommunicative, smug, languid, slow
I Demanding, impatient, dependent, immature, imaginative, introspective, kindly, gentle, aesthetically fastidious, frivolous, attention-getting	vs. Emotionally mature, independent-minded, set and smug, hard, cynical, lacking artistic feeling, responsible, self-sufficient

Guilford used a different approach to the problem, administering several personality tests to large numbers of subjects, computing hundreds of intercorrelations between pairs of scores for the same individuals, and completing a factor analysis of the correlation matrix. As a result of this research, Guilford tentatively identified ten relatively independent personality traits:

- G General activity: hurrying, liking for speed, liveliness, vitality, production, efficiency
- R Restraint: serious, deliberate, persistent, vs. carefree, impulsive, excitement-loving
- A Ascendancy: self-defense, leadership, bluffing, speaking in public, vs. submissiveness and hesitation

² Raymond B. Cattell, *Description and Measurement of Personality* (New York: Harcourt, Brace & World, Inc., 1946).

- S Sociability: many friends, seeking friends and social activities, seeking limelight vs. few friends, shyness
- E Emotional stability: evenness of moods, optimistic, composure, vs. fluctuation of moods, pessimism, daydreaming, excitability, feelings of guilt, worry, loneliness, and ill health
- O Objectivity: thick-skinned, accurate, observing, vs. hypersensitive, self-centered, suspicious, having ideas of reference
- F Friendliness: tact, acceptance of domination, respect for others, vs. hostility, resentment, desire to dominate, and contempt for others
- T Thoughtfulness: reflective, observing of self and others, mental poise, vs. interest in overt activity and mental disconcertedness
- P Personal relations: tolerance of people, faith in social institutions, vs. fault-finding, uncooperative, suspicious, self-pitying
- M Masculinity: interested in masculine activities, not easily disgusted, hard-boiled, inhibits emotional expression, little interest in clothes and styles, vs. easily disgusted, fearful, romantic, emotionally expressive, and dislike of vermin³

Some of the personality traits that had been hypothesized by the pioneers in personality testing, such as dominance-submission and emotional stability, were confirmed by these studies as independent personality traits. Others, such as extroversion-introversion, did not emerge as unitary traits.

The Limitations and Advantages of the Transparent Personality Inventory

If an individual has sought guidance, use of a personality questionnaire with direct, transparent questions may be advisable. The early inventories were quite transparent, while recently published inventories have incorporated a number of devices to minimize the probability that the subject is answering the inventory so as to create a good impression. Actually, whether one uses a more transparent or more subtle approach depends largely on the situation, the attitude of the examinee, and the use to be made of the results. The fact that the more transparent personality inventories can be faked does not imply that they always are. According to one research study, if students are convinced that their replies will be kept confidential and used as a basis for helping them, they will usually give frank responses to personality questionnaires.⁴

If a person voluntarily indicates symptoms of poor mental health through

³ Edward B. Greene, *Measurements of Human Behavior*, rev. ed. (New York: The Odyssey Press, Inc., 1952), pp. 636-638.

⁴ Dora E. Damrin, "A Study of the Truthfulness with Which High School Girls Answer Personality Tests of the Questionnaire Type," *Journal of Educational Psychology*, vol. 38 (April 1947), pp. 223-231.

his replies to direct questions, his replies are easy to interpret and counseling can proceed more effectively. An individual who has sought help is usually motivated to answer truthfully; he has faced up to the fact that a frank discussion of his problems may be a necessary prelude to improvement.

If a personality inventory is to be used as an aid in *selection and classification*, the results from a transparent inventory may be useless, or even misleading. A more indirect approach must be used. By using subtle questions, or by employing forced-choice questions, we may "trap" the individual into revealing more about himself than he intended to do. However, the more indirect the questions, the less confident we can be of our interpretation of his scores. Using indirect questions, or the forced-choice technique, makes an inventory more difficult to fake; hence inventories that are less transparent are preferred in prediction situations. However, as we gain in fake-resistance and predictive validity, we may lose in meaningfulness or construct validity.

Rimland found that warning examinees that the inventories would be scored for truthfulness brought good results. That is, he compared the extent of faking under standard instructions and under instructions to warn students that a "lie" score would be obtained. He found that informing examinees about validation scores reduced faking to a minimum.⁵

The problem of "faking," however, is not the only one we face in using personality inventories. Just as important as frankness is the student's *insight* into his own behavior—his ability to describe his own reactions without distortion.

Adjustment is an emotional matter, something at which people cannot look in the light of pure reason alone. In contemplating their own adjustment, they are more likely to become biased, prejudiced, secretive, and deceitful of others and of self, than when contemplating their achievement in geometry, their physical health, or even their mental ability.⁶

It has been found that the maladjusted individuals are the ones who are most likely to distort their responses on personality inventories.⁷ Hence, it is most difficult to obtain valid responses from those students for whom

⁵ Bernard Rimland, *The Development of a Test for Selecting Career Motivated NROTC Applicants*. Bureau of Naval Personnel Technical Bulletin 57-58, 1957.

⁶ H. H. Remmers and N. L. Gage, *Educational Measurement and Evaluation* (New York: Harper & Row, Publishers, Inc., 1943), p. 338.

⁷ P. E. Vernon, "Review of Humm-Wadsworth Temperament Scale," *The 1940 Mental Measurements Yearbook* (Highland Park, N.J.: The Gryphon Press, 1941), pp. 122-124.

diagnosis is most important. Although there is little doubt that a student who makes a *low* score on a personality questionnaire should receive further study, there is no guarantee that *high* scores indicate good adjustment. The maladjusted student may be untruthful about his feelings and actions or may have built up defense mechanisms that obscure his insight into his own problems.

Of course, the same factors that make it difficult to obtain frank and insightful responses on a personality questionnaire apply also to such techniques as interviewing student or parent and observing student behavior. As they grow older, people learn in varying degrees to conceal feelings and attitudes that are not socially approved.

Attempts to Increase the Validity of Inventory Results

The transparent inventory provides only a summary record of those symptoms and self-criticisms the individual is willing to check. When an individual seeks guidance, a transparent inventory can facilitate the counseling process by indicating the problems he recognizes and may be willing to discuss. For use in vocational guidance, diagnosis of maladjustment, and personality research, however, the transparent inventory has proved quite unsatisfactory.

Ellis, after reviewing studies reported in the literature, concluded that "group-administered paper-and-pencil personality questionnaires are of dubious value in distinguishing between groups of adjusted and maladjusted individuals and that they are of much less value in the diagnosis of individual adjustment or personality traits."⁸

Some of the more recently published inventories have incorporated features, such as the use of forced-choice questions⁹ (which have made them more fake-resistant) and verification scores (which have helped test-users to identify inventories in which students have distorted their responses to create a good impression). Even with these improvements, however, routine use of personality inventories seems inadvisable. Adequate psychological

⁸ Albert Ellis, "The Validity of Personality Questionnaires," *Psychological Bulletin*, vol. 43 (September 1946), p. 426.

⁹ Forced-choice questions are of the type used in the *Kuder Preference Record* in which the examinee is forced to choose between responses. In forced-choice inventories, the responses among which a student must choose are first matched with respect to their *social desirability*. In such an inventory, the examinee is prevented from choosing most of the socially desirable items and rejecting the socially undesirable ones. For a more adequate explanation of the response tendency to answer questions in a socially desirable manner, see Allen L. Edwards, *The Social-Desirability Variable in Personality Assessment and Research* (New York: Holt, Rinehart and Winston, Inc., 1957).

cal background and experience are required for drawing sound inferences from personality inventory results. Hence, such inventories are best used as one approach in the study of selected students by qualified psychologists and counselors.¹⁰ The cumulated research on the inadequate concurrent and construct validity of personality inventories has led to many attempts to improve their validity.

DISGUIISING INVENTORY ITEMS Some test authors have attempted to *disguise* their test items, anticipating, in a sense, the student's tendency to rationalize his behavior. The following examples of partially disguised, or subtle, questions are taken from the *California Test of Personality*:

Are your tests so hard and unfair that it is right to cheat?
Do your classmates quarrel with you a great deal?
Do you suffer more than most people when you are ill?

Wiener¹¹ contends that partially disguised or subtle items are best for personality inventories used with normal subjects who frequently show a tendency to respond in a socially desirable way, while obvious items function best for abnormal subjects, especially if the latter frankly acknowledge their need to report symptoms and gain help thereby. Cronbach¹² has suggested that interest and personality inventories be scored separately for obvious and subtle items.

OBSCURING THE SCORING PATTERN Authors have also tried to reduce the examinee's awareness of the traits or components being measured by arranging the items pertaining to each component (such as antisocial behavior or nervous mannerisms) in random order throughout the test. When similar items are grouped together, the inventory becomes more transparent and the examinee can more readily respond to items so as to create any desired impression. Scattering items for each trait throughout the inventory complicates hand scoring but probably increases the validity of test scores, as compared with the procedure of grouping together all questions on antisocial behavior, all questions on nervous mannerisms, and the like.

¹⁰ As will be explained later, a problems checklist can be administered to large numbers of students as an aid in discovering common problems for consideration in group guidance work or for identifying students who voluntarily admit problems they consciously face and on which they would like individual counseling.

¹¹ Daniel N. Wiener, "Subtle and Obvious Keys for the Minnesota Multiphasic Personality Inventory," *Journal of Consulting Psychology*, vol. 12 (May-June 1948), pp. 164-170.

¹² Lee J. Cronbach, *Essentials of Psychological Testing* (New York: Harper & Row, Publishers, Inc., 1960), p. 458.

USING VERIFICATION SCORES TO IDENTIFY INVALID INVENTORIES Considerable work has also been done on the development of sets of items that can be scattered through an inventory and scored so as to reveal test-taking attitudes. Verification or validation scores, as they are called, are especially needed if fairly transparent questions are being used in situations in which many examinees would feel on the defensive. The more threatened the examinee feels by the test situation and his own feelings of inadequacy, the more likely he is to make "good impression" or "socially desirable" responses. The more transparent the inventory, the easier it is for the examinee to "fake good" or present a façade. The *California Psychological Inventory* includes a number of items selected because examinees tend to respond differently to them under standard and "fake good" directions. These items constitute a "good impression scale."

Several inventories include a number of questions on common faults and frailties to which the majority of persons make the unfavorable response; for example, "Have you ever pretended to know something you did not know?" A person who answers a large number of these questions in the favorable direction has too high a "lie" score to have his inventory considered valid.

The *Minnesota Multiphasic Personality Inventory* (MMPI), a very extensively studied inventory used chiefly by clinical psychologists, has several validation scores: a lie score, a K-score of test-taking defensiveness (similar to a good-impression scale), an F-score on deviant or rare responses, a question score (the total number of items to which the examinee responds with "cannot say"), and an inconsistency score. When several validation scores are available, they can be used in combination. For example, a person with high scores in both deviancy and inconsistency has probably filled out the questionnaire very carelessly; while a person with a high deviancy score and high consistency may actually show bizarre responses in real-life situations.

USING CORRECTION SCORES Most types of validation scores merely enable the psychologist or counselor to discard inventories that are suspected of having been filled out carelessly, or to create a good impression. However, some validation scores, for example, the K-score on the MMPI, are used to *correct* for the tendency to "fake good" or respond defensively. If an examinee has a high K-score (which presumably reflects defensiveness in test-taking), his scores are *corrected* for this tendency. His corrected inventory is salvaged rather than discarded.

SELECTING AND KEYING ITEMS ON THE BASIS OF EMPIRICAL DATA Another approach to increasing inventory validity is to select and key items on the basis of research findings regarding the typical responses of criterion

groups. The student will recall that the responses to items of Strong's interest inventories were assigned weights on different occupational keys in terms of the differential responses of occupational groups, rather than on any logical, *a priori* basis. The MMPI is an example of a personality inventory in which the items are empirically selected and keyed. Items that paranoid patients answered differently from normals were assigned to the paranoid key. Since paranoids tend to insist on their own ideas, the following question would *logically* be assigned to that key: "It takes a lot of argument to convince some people of the truth." However, since paranoids did not check this statement as true any more frequently than did normals, it is *not* included in that scale. Answering false to this question, however, was found to be associated with hysteria and is therefore keyed on the hysteria scale.

Responses to questions were treated as symptoms or signs of a diagnostic category in the "concurrent validity" sense; they were not considered as a representative sampling of a defined area as in the "content validity" sense. As the reader can see, the empirical approach makes no assumptions regarding the examinee's honesty or insight. Rather, it evaluates each item empirically in terms of its relationship to the criterion, for example, to membership in a diagnostic group or (for trait inventories for normal subjects) with ratings by close associates on the trait presumably being measured. The reader will recognize, however, that the empirical basis for selecting items is no royal road to the development of adequate personality inventories because it is so difficult to obtain adequate criterion data.

On personality questionnaires designed to identify *tendencies to abnormal behavior*, the items are usually validated by comparing the responses of subjects in various diagnostic categories in mental institutions with those of so-called normal subjects. Those questions are included that tend to be answered differently by normal and abnormal subjects. The criterion for item validity is membership in a normal group or in a diagnosed group of patients from a mental hospital or clinic. Appropriate methods of developing and validating inventories designed to have concurrent validity are presented in Table 4.3.

If a personality inventory is being used in the selection of employees,¹³ items will be selected and keyed in terms of their *predictive validity* for some such criterion as job turnover or job production (for example,

¹³ "Standards of Ethical Behavior for Psychologists," *American Psychologist*, vol. 13 (June 1958), pp. 266-272. When personality inventories are used as a basis for selection and classification, it is essential that the psychologist make sure that examinees understand the purpose for testing and the way in which the results will be used. The "Standards of Ethical Behavior for Psychologists" must be scrupulously observed.

amount of insurance sold). Such predictive validity studies are best done in each company; job requirements, working conditions, and criteria of success vary enough that items that work best in one situation may not be the ones that have highest predictive validity in another. The item validities obtained in one group of employees need to be checked or cross-validated with another group; and after a period of years, restudy is again needed to see if the predictive value of items has changed.

When attempts are made to measure *personality traits of normal individuals*, such as dominance or sociability, questionnaire items are often selected by comparing the responses of students who are rated high by teachers or psychologists in the personality trait being measured with the responses of those who are rated low in the same trait. Correlation with ratings is a questionable method of validating items because the criterion ratings tend to be unreliable, to be affected by extraneous factors (such as the student's attitude toward school and the teacher), and to have no demonstrated construct validity of their own.

Another basis for item selection is to compare the responses to individual questions of students who made high *total* scores in dominance, sociability, and the like, with the responses of students who make low total scores on these same traits on the preliminary edition of the test. This technique of item selection will make the subtests more homogeneous but ignores the problem of whether the subject's verbal responses are related to his behavior or other people's impressions of him. No external criterion is involved. Methods of developing and validating tests to be used in trait description involve many approaches, summarized in Table 4.9.

The Reliability of Personality Inventories

Measuring test reliability by the test-retest method is inappropriate for personality inventories. The inconsistency in a student's responses to questions in a personality test may actually reflect an important aspect of his personality. Cattell found, for example, that individuals who changed their responses the most when tests of interests and attitudes were readministered after a few days tended to rate high on emotional instability, one of the most important source traits in personality. Because a test-retest reliability coefficient would be affected by individual differences with respect to this trait, internal consistency methods are preferred in checking the reliability of personality inventories. The split-halves method and the Kuder-Richardson method, which reveal only the internal consistency of the test, are most frequently used.

The reliability coefficients of personality inventories tend to be considerably lower than those for ability tests of the same length. One reason is

that many responses in the area of personal-social adjustment tend to be specific to the situation. A person who might show extrovertive behavior on a job in which he had high competence might be introvertive in social situations. Even within the realm of social situations, there will be inconsistencies as the person is in small or large groups, with persons of the same or opposite sex, or with his peers as compared with those in a position of authority.

In interpreting personality profiles, it is especially important to check the test manual concerning the reliability of subtests and the intercorrelations between them. When one interprets such data with the aid of Table 3.8, one may find that the difference scores have such low reliability that only the largest differences can be used as a basis for making inferences about intraindividual differences for a student.

Interpreting Results from Personality Inventories

In the typical personality inventory, relevance has been sacrificed in some degree in order to achieve objectivity of scoring. For example, more valid and significant information would undoubtedly be obtained if, instead of asking "Do you daydream frequently?" one asked a student about his daydreams or the circumstances that led him to have "spells of the blues." Such questions, however, would not ordinarily be included in a personality inventory, both because of the difficulty of obtaining frank replies and because of the subjectivity that would be involved in scoring and interpretation.

In a personality inventory, the extent of a student's maladjustment is presumably indicated by the *number* of his unfavorable responses. The questionnaire does not reveal the intensity of feeling involved or the appropriateness of the behavior in terms of environmental factors. For example, if a student answers "Yes" in response to the question, "Do you get excited when things go wrong?" it would be necessary to obtain additional information (through observation or through interviews with the student and his parents) before appraising this response in terms of *degree* of maladjustment and type of assistance needed. Through the use of other techniques, one could find the answers to such questions as the following:

How excited does he get?

Is the degree of excitement unusual for a person of his age?

Is his excitement disproportionate to the stimulus situation?

Is his excitement so great that his behavior becomes disorganized?

Is he overexcitable only when fatigued?

About what things does he become excited?

In what way does he manifest his excitement—in physical aggressiveness, in blaming others, in quiet tenseness, or in tearful self-pity?

In other words, personality questionnaires can aid in the first level of diagnosis—identifying those who need help.¹⁴ They can assist also in the second level of diagnosis—locating areas of difficulty—especially when responses to individual items are studied. For the third level of diagnosis, however—identifying causative factors—it is especially important that the *leads* given by the personality inventories be used as a basis for further study of the student through the combined use of observation, student interview, parent conference, and other techniques presented in this and the succeeding chapter.

Students' replies to individual items of a personality questionnaire may provide leads for observation and for follow-up interviews. In fact, scanning replies to individual questions may prove to be more valuable than a routine analysis of test scores. If one suspects, for example, that a student's school adjustment difficulties are rooted in his family relationships, he may wish to examine the student's replies to those questions involving parent-child relationships, such as:

- Are you scolded for many little things that do not amount to much?
- Do you feel that you are bossed too much by your folks?
- Have things ever been so bad at home that you have had to run away?
- Do you wish that more affection were shown by more members of your family?
- Do your folks appear to doubt whether you will be successful?¹⁵

The study of student replies to such groups of related questions may provide extremely helpful leads for further study. In his conferences with parents, however, the counselor should avoid any reference to specific statements made by a student or to the inventory as a source of information. It is both unnecessary and unwise, for example, for a counselor to reveal that his interest in "problems you may have with John at home" grows out of John's low test score in "family relationships." Nor should personality inventory scores be cited in discussing the student's problems with his parents. Students' replies to items of a personality inventory should be considered confidential information.

Unless personality inventories have been developed according to the methods recommended in our discussion of "construct validity" in Chapter 4, we are *not* justified in interpreting the personality inventory scores as representing a student's status with respect to underlying personality dimension or traits. We should, rather, view a student's personality inven-

¹⁴ The three levels of diagnosis are explained in Chapter 14.

¹⁵ *California Test of Personality, Secondary* (Monterey, Calif.: California Test Bureau, 1953).

tory as a summary of the symptoms and self-criticisms he was willing to report on a specific questionnaire administered on a specific occasion. The student's results on another questionnaire with the same trait label may give quite dissimilar results. As Cronbach emphasizes

Trait names . . . are a source of serious confusion in the personality field. The meaning of "introvert" . . . represents for one author a brooding neurotic, for another anyone who would rather be a clerk than a carnival barker. "Ascendancy" ranges from spontaneous social responsiveness, in one theory, to inconsiderate and overbearing behavior in another. . . . In the present Babel of trait names, the only useful way to discuss personality test data is to speak of "Guilford's Ascendancy score," "CPI Dominance score," according to the measure used.¹⁶

Although it is neither feasible nor desirable to discuss many of the published inventories that are now available, a few of the widely used inventories will be briefly discussed as illustrative of the major types of inventories available.

INTERPRETATION OF THE CPT, A TRANSPARENT INVENTORY Illustrative of a transparent inventory of personality "traits" is the *California Personality Test* (CPT), which has forms available for use at all grade levels and with adults. Although an attempt has been made to disguise questions, this inventory is transparent to the test-wise student. The fact that items of a type are grouped together for ease in scoring makes it evident that one is being asked about nervous symptoms, withdrawing tendencies, and the like. Hence, the use of the CPT should be restricted to situations in which we have every reason to believe that frank responses will be given. The reliability of difference scores should be noted in interpreting CPT profiles.

Another problem that complicates the interpretation of transparent personality inventories is that the well-adjusted person, especially the well-adjusted adult, is more likely to admit his faults than the person who is more insecure and has greater need to distort his self-appraisals. Loevinger¹⁷ has pointed out that psychologically mature college students and adults may get lower scores than persons whose self-concepts are at a less mature, overconforming stage of development. One should probably be suspicious of the consistently high personality profile on a transparent inventory. Even

¹⁶ Lee J. Cronbach, *Essentials of Psychological Testing* (New York: Harper & Row, Publishers, Inc., 1960), pp. 467-468.

¹⁷ Jane Loevinger, "A Theory of Test Response," *Proceedings of the 1958 International Conference on Problems of Testing* (Princeton, N. J.: Educational Testing Service, 1959).

though the examinee is telling the truth as he sees it, he may have a special need to see himself as having all the characteristics expected of him by society.

INTERPRETATION OF THE MMPI AND ITS ADAPTATIONS (EMPIRICALLY-KEYED INVENTORIES) Although the items of the MMPI inventory were originally selected and keyed in terms of their relationship to diagnostic categories (depression, hysteria, and the like), it is now recognized that the results are best interpreted in terms of cumulated research data on subjects with various *coded profiles*, that is, various combinations of high and low scores on the original scales. The interpretation of the MMPI requires extensive psychological background, supervised experience in its use, and thorough familiarity with such aids as are listed in the chapter bibliography and with more recent research studies. Approximately one hundred new research studies on the MMPI appear each year. Even when the currently approved methods of interpretation are used by highly qualified psychologists, the scores should be used as a basis for *hypotheses* to be checked by other methods, rather than conclusions. When used in this way, the MMPI seems to have considerable value for clinical psychologists.

Two adaptations of the MMPI that are more suitable for use with normal subjects are the *California Psychological Inventory* (CPI) and the *Minnesota Counseling Inventory* (MCI). Both tests have verification scores. Many research studies have been completed on the relationship of scores on these inventories and such significant variables as underachievement, delinquency proneness, and others. Bibliographies of such studies can be obtained from the publishers.

On any personality inventories, one should study combinations of scores on different subtests, rather than single subtest scores. For example, Gough, author of the CPI, suggests that when a high score on Ai (achievement through independence) is accompanied by a high score on Ac (achievement through conformity), the person is likely to be efficient, well organized, and stable; whereas a high score in Ai accompanied by a low score in Ac tends to be found in people who are demanding and dominant.¹⁸

INTERPRETATION OF THE EPPS (A FORCED-CHOICE INVENTORY) DESIGNED TO MEASURE CONSTRUCTS Another inventory on which considerable research data are available is the *Edwards Personal Preference Schedule*. This inventory differs from those previously discussed in at least two ways: (1) its design grows out of psychological theory (that is, the

¹⁸ Harrison C. Gough, *Manual, California Psychological Inventory* (Palo Alto, Calif.: Consulting Psychologists Press, 1957).

Murray¹⁹ theory of psychological needs); and (2) it utilizes the forced-choice method, the alternative responses being matched with respect to their rated social desirability. The first of these two characteristics implies that this inventory must be evaluated in terms of its construct validity as a measure of hypothesized dimensions. The second characteristic indicates that the inventory is designed to be fake-resistant; that is, the examinee is unable consistently to choose socially desirable responses.

Although this type of inventory has many advantages, it has a few disadvantages: (1) the forced-choice method shows only intraindividual differences, not the relative strength of a trait with respect to other examinees; (2) unless students are highly motivated in self-appraisal, they tend to resist this type of test because they must make difficult choices and they feel frustrated in their attempt to paint a desirable self-picture; and (3) this type of test tends to have low reliability coefficients for subtests, partly because of the fact that the student experiences such uncertainty in choosing that he is likely to change many choices on retesting. The first characteristic was considered in the discussion of the Kuder inventories in the preceding chapter. The second problem can be met, in part, by interesting the student in self-appraisal. In connection with the third, we have to face the fact that the higher reliability of many of the other inventories with which we might compare the EPPS is attributable in large measure to the consistency with which the examinee can communicate the impression he wishes to make on a transparent inventory in which his choices are not limited by the preference or forced-choice approach.

INTERPRETATION OF PROBLEMS CHECKLISTS A quite different type of personality questionnaire is the problems checklist. The items on such a list are intended to be a representative sampling of problems in different areas. No claims are made that the checklist measures personality traits or that its use can entice the student into revealing any problems he does not choose to report.

An illustration of a problems checklist summarized by *areas of adjustment* is the *Mooney Problem Check Lists*,²⁰ one form of which can be used as early as the seventh grade. Student problems are classified under the following areas: Health and Physical Development, School, Home and Family, Boy-Girl Relations, Relations to People in General, and Self-centered Concerns.

Another example is the *SRA Youth Inventory*, developed for use in

¹⁹ Henry A. Murray and others, *Explorations in Personality* (New York: Oxford University Press, 1938).

²⁰ Ross L. Mooney, *Problem Check Lists* (Columbus, O.: Ohio State University, 1943).

grades 7-12, but with separate profiles for the junior and senior high school levels. The leaflet on which the student draws his profile provides three and one-half pages of discussion designed to help him understand and use the inventory results. A description of the types of problems in each area, suggestions on how to get help in solving problems, and a concrete example based on one student's experience are included. The problems reported by students are summarized under eight areas, as follows: My School, Looking Ahead, About Myself, Getting Along with Others, My Home and Family, Boy Meets Girl, Health, and Things in General. Item norms for different grade and sex groups are given in the manual. In addition, a *Basic Difficulty Key* is supplied for use by the counselor in indicating problems that may be caused by more serious personality difficulties. The *SRA Junior Inventory* for grades 4-8 is also available.

The more recently published *Billett-Starr Youth Problems Inventory* has separate forms for grades 7-9 and 10-12. Like the others, this inventory is designed for screening those students who need and wish individual counseling, as well as for identifying common problems that might be approached through group guidance procedures. Totaling responses for area scores is *not* recommended.

Students who voluntarily admit on a problems checklist that they have problems they would like to discuss can be called in for counseling interviews. Use of such a checklist helps the counseling staff to make effective use of the limited time available. One must recognize in interpreting problem checklists, however, that only consciously felt problems that the student is willing to report will be checked. The results should be interpreted as suggesting areas worthy of exploration. Such checklists can provide a good starting point for individual or group guidance. Since there has been considerable parent criticism of widespread administration of personality inventories of any type, it is probably desirable that an advisory committee of representative parents assist in the planning of projects in which such checklists are routinely administered and in the interpretation of such projects to other parents.

PROJECTIVE TECHNIQUES

Clinical psychologists have tended to criticize the kinds of personality descriptions provided by personality inventories. These critics prefer approaches to personality study that consider the individual as a whole and that try to explain why people behave as they do, not merely what they are like. They contend that one cannot adequately describe complex human personality by summing a series of trait scores. As Murphy says:

There are many different ways in which the simplest traits of the individual may be put together; some operate summatively, others subtractively. Sometimes when one trait is present it acts almost like an enzyme, allowing the more effective utilization of another trait.

Recently biologists . . . and other originators of "general system theory" have shown us that wherever living systems are involved, the problem of organization and emergence takes over and complicates the problem of showing how the individual trait reflects the system of which it is a part. It does not shock us today to be told that a patch of red in a landscape will look differently if the context is altered. But it still does seem to bother us . . . to be told that a trait is operationally different when it appears in different contexts. Coolness, for example, or the maintenance of a low level of affect, is a very different thing in a danger situation and in a social gathering.²¹

Projective Techniques Used by Clinical Psychologists

Anastasi²² has classified projective techniques into five major categories:

1. Associative techniques, in which the individual responds to a stimulus by giving the first reaction that occurs to him. This approach, initiated by Galton and Jung, has been extensively used in screening subjects for psychiatric study. One of the most widely used lists of stimulus words is the *Kent-Rosanoff Free Association Test*. On this test, the psychologist can check the subject's responses to each stimulus word to see the frequency with which that response was given in a standardization group of 1,000 normal adults.

One of the oldest and most widely studied of projective techniques, *The Rorschach Ink Blot Test*, is classifiable in this first group as an associative technique. The test is administered individually. Although group forms are available, they are not considered to be comparable. In the administration of the Rorschach, the subject indicates orally what he thinks each ink blot might be, and the examiner records his responses. During an inquiry that follows the initial test, further responses are sought; and the psychologist "tests the limits" to see if the subject is capable of giving certain types of responses that he has not previously given.

Psychologists differ widely in their appraisal of the value and limitations of the Rorschach. Although more than two thousand publications on the Rorschach are available, there are few well-designed studies concerning the validity of the assumptions involved in its interpretation. Several reviews of the test are given in the *Fifth Mental Measurements Yearbook*, together with an extensive bibliography of research studies.

The recently developed *Holtzman Ink Blot Technique* claims to provide

²¹ Gardner Murphy, "Concepts of Personality—Then and Now," 1956 *Invitational Conference on Testing Problems* (Princeton, N.J.: Educational Testing Service, 1957), pp. 43–44.

²² Anne Anastasi, *Psychological Testing* (New York: The Macmillan Company, 1961), p. 566.

the advantages of the Rorschach in clinical work and yet overcome some of its disadvantages. The number of ink blots is much larger; instead of 10 ink blots, there are 90 (organized into two parallel forms of 45 each). Since only one response is obtained for each ink blot, the total number of responses is comparable from subject to subject. Computer analysis of the responses of hundreds of subjects has produced a scoring guide that contributes to high interscorer reliability. Percentile norms on 22 response variables have been prepared for eight groups, ranging in age from five-year-olds to adults, and including defined clinical groups. Although the authors have reported in detail concerning the development and standardization of this test,²³ years of evaluated experience in its use will be required before its contribution to the appraisal of personality can be properly assessed.

2. Construction procedures, which require the subject to create or construct a product, such as a story. The tasks are usually introduced as a test of imagination or creative ability, and interpretation of the results typically involves a content analysis of the story or other product.

The best-known test of this type is the *Thematic Apperception Test* (TAT), which has been extensively used in both personality research and clinical work. The subject is asked to tell a story about each of several ambiguous pictures, selected because of their relationship to characteristic conflicts, personality needs, and environmental pressures. Although several quantitative scoring plans have been developed and used in personality research, clinical psychologists tend to interpret the content in terms of their own perception of recurrent underlying themes and in terms of other information they have about the subject.

3. Completion tasks, such as completing sentences or incomplete reaction stories. The "incomplete sentence" technique has been used and studied for years. In 1950, a selected list was published in a more objective test form as the *Rotter Incomplete Sentences Test*.²⁴ Another widely used list was prepared by Rohde in 1957.²⁵

Another completion task, widely used in personality research and clinical work, is the *Rosenzweig Picture-Frustration Study*, in which the subject reacts to a series of cartoonlike drawings. In each drawing two characters are involved in a mildly frustrating situation of a type that frequently occurs in everyday life. The subject is asked to indicate what the frustrated person is probably saying. On the assumption that the subject identifies with the frustrated individuals, his responses are classified according to whether they represent (1) "extrapunitive" responses (aggression directed outward), (2) "intropunitive" responses (aggression turned in upon the subject himself), or (3) "impunitive" (attempts at glossing over or evading the situation). Only moderate agreement among scorers on the classification of subjects' verbal responses has been obtained.

4. Choice or ordering devices, calling for the rearrangement of stimuli, the

²³ Wayne H. Holtzman, and others, *Inkblot Perception and Personality* (New York: The Psychological Corporation, 1961).

²⁴ J. B. Rotter and Janet E. Rafferty, *Manual for the Rotter Incomplete Sentences Blank* (New York: The Psychological Corporation, 1950).

²⁵ Amanda R. Rohde, *The Sentence Completion Method* (New York: The Ronald Press, 1957).

recording of preferences, and the like. Since these tasks require fairly simple responses from the subject, the scoring can be entirely objective, although it may be time-consuming. One of the oldest and best-known tests of this type is the *Szondi Test*.²⁶ Photographs of patients with various types of mental illness are presented to the subject, who indicates which photographs he prefers. It is assumed that the subject will tend to choose photographs of patients with tendencies similar to his own. Validity studies have been disappointing. The *Tomkins-Horn Picture Arrangement Test* (PAT) requires that the subject react to each series of three pictures, arrange the three in the order "which makes the best sense" to him, and write a sentence for each of the three pictures so as to tell the story he has in mind. The test can be administered in groups and objectively scored. Temporal stability of scores is low, and adequate validity studies have not been completed. Like many personality tests, the PAT requires much more research to aid in meaningful interpretation of scores.

5. Expressive methods, which differ from construction procedures in that the individual's style or method is evaluated as well as his product. Almost every technique and type of subject-matter have been used. One of the best-known examples is the *Draw-a-Person Test* by Machover.²⁷ The subject is asked to "draw a person"; while he draws, his sequence in drawing and time used are recorded, as well as his comments and questions. When he completes the drawing, he is asked to draw a person of the opposite sex from the one he chose for his first picture. The results are interpreted qualitatively, a composite personality description being prepared by the examiner from an analysis of its special characteristics.

The use of puppets, dolls, and miniature objects in original dramatizations would also be included under this heading. Techniques of play therapy have been adapted for use in projective testing, the examiner noting the objects the child selects, how he uses them, his verbalizations, and other behavior as he acts out his feelings. The *Driscoll Play Kit*, for example, illustrates the type of materials available to clinical psychologists. The kit opens to form an apartment inhabited by five plastic dolls with movable joints (designed to represent mother, father, brother, sister, and baby).

Projective techniques tend to afford wide bandwidth and low fidelity. Hence, they are best used as exploratory techniques in the early phases of clinical study. They are best interpreted in combination with data from other sources that can serve to confirm or question hypotheses developed. They have the advantage of being highly interesting to most subjects and highly fake-resistant. Some of them can be used with children or adults who have difficulty in communicating verbally. Some involve nonverbal communication, which is not subject to as much restraint or censorship as verbal communication. Those that do depend on verbal communication, such as the picture-story tests, utilize materials and are administered in a

²⁶ Susan K. Deri, *The Szondi Test* (New York: Grune and Stratton, 1949).

²⁷ Karen Machover, *Personality Projection* (Springfield, Ill.: Charles C Thomas, 1949).

setting that encourages the subject to reveal fantasy material which he might ordinarily suppress.

Separate volumes have been written about each of the various techniques listed above. Courses in projective techniques, which require prerequisite education in clinical psychology, must be supplemented by supervised experience in the administering, scoring, and interpretation of these techniques.

TEACHER USE OF ADAPTATIONS OF PROJECTIVE TECHNIQUES Of the techniques classifiable under projective techniques, teachers and counselors may be able to use to advantage: reaction stories, open questions, incomplete sentences, and study of students' creative writing and art products.

In using open questions or incomplete sentences, teachers or counselors may use them individually with students on whom case studies are being made in order to understand their problems. Administered in privacy to a student with whom rapport has been established, they may provide leads for further study. If open questions or incomplete sentences are to be administered to a class, they should be of a more impersonal type (for example, "My most difficult subject . . .," "My favorite recreation . . .,") or should allow the individual considerable latitude in his choice of subjects (for example, "When I grow up . . .," or "If I had three wishes, . . ."). Because of the understandable reluctance of students to make negative statements in a classroom situation, it may be advisable to phrase most items positively, for example, "What I like about my home," or "What others have said they like about me."

Teachers need to be especially careful in making inferences from students' creative writing and their art products. Unless a theme utilized in a student's story is repeated in the same or similar form in other writings, it may represent only a plot from a television drama recently seen or a story recently read.

Although psychologists are agreed that children express their emotions and conflicts through art,²⁸ they are not ready to present a list of principles that teachers can use with confidence in the interpretation of children's art work. The art work of all children is not equally revealing. Some children, especially in the middle and upper grades of elementary school, become so engrossed in the problem of representing objects and people that their work is no longer an emotional outlet for them.²⁹

Just as the teacher must not attach too much diagnostic significance to a single story, she must be similarly cautious about the interpretation of

²⁸ Florence L. Goodenough and Dale B. Harris, "Studies in the Psychology of Children's Drawings, 1928-1949," *Psychological Bulletin*, vol. 47 (September 1950), pp. 369-433.

²⁹ *Ibid.*, p. 10.

isolated paintings. A *series* of paintings will reveal a child's characteristic style, colors most frequently used, recurring content, maturity in representation, characteristic distortions or omissions in the human figure, and the like.

Since expression of emotions through art is so highly individualized, it is obvious that norms for interpretation are difficult to develop. The significance of certain colors and techniques varies with the maturity of the child. For example, Alschuler and Hattwick found that consistent emphasis on cold colors was not typical of the happy nursery-school child, and that children of these ages who consistently favored cold colors showed overcontrolled behavior. Among older children, however, preference for the cooler colors was no longer indicative of poor adjustment.³⁰

SUMMARY STATEMENT

Personality inventories and projective techniques are alike in that their use by unqualified persons is hazardous; they are widely different in that they represent the two extremes of the psychometric vs. the clinical approach to the study of personality. Moreover, personality inventories are structured self-report questionnaires designed to minimize ambiguity; while projective techniques involve observing the examinees' behavior in situations that are designed to be unstructured or ambiguous and in which examinees are encouraged to react in a highly individual manner. In general, psychologists who make considerable use of one of these approaches tend to distrust the other.

As psychologists have attempted to develop inventories that would provide personality profiles for individuals, each author has tended to design and score his inventories on the basis of his own list of personality traits. More recently, the techniques of factor analysis have been applied in an attempt to identify independent personality traits. The results of research studies by Cattell and Guilford were found to show considerable agreement.

The chief type of personality test used in schools is the personality inventory or questionnaire, in which students answer a series of questions concerning their attitudes, feelings, and behavior. The validity of personality-inventory results depends not only on the care with which the questions are selected but also on the ability and willingness of students to give truthful responses. Attempts to check on the reliability of personality inventories and other appraisal techniques are complicated by the fact that the very inconsistency in a student's response may actually reflect an attribute of his personality.

Some personality inventories are quite transparent; the examinee can easily discern the traits the test author is trying to measure. Other inventories have been carefully designed to disguise the dimensions being measured, and to thwart the efforts of the examinee who wishes to create a desirable impression. The relative merits and limitations of the transparent inventory are considered,

³⁰ Rose H. Alschuler and L. W. Hattwick, *Painting and Personality, A Study of Young Children* (Chicago: University of Chicago Press, 1947), p. 17.

as well as the many techniques now being used in an attempt to increase the validity of inventory results.

In the section on interpretation of personality inventories, separate consideration was given to the interpretation of results from (1) a transparent inventory, (2) an empirically keyed inventory, (3) a forced-choice inventory designed to measure constructs, and (4) problem checklists. Careful consideration of this material will help the reader to realize that the best inventory for one purpose will be quite unsatisfactory for another, and that each type has to be interpreted in terms of the type of inferences justified by the approach used in test construction, as well as the situation in which the test was administered (or the examinee's perception of that situation).

Projective tests may encourage individuals to give their spontaneous reactions to ambiguous stimuli such as ink blots or to create individualized products such as stories about pictures involving conflict situations, which permit of a variety of interpretations. In some tests the individual is asked to complete unfinished sentences or stories; in others he indicates his preferences, for example, by arranging stimuli in a preferred order. Still other tests involve an evaluation of the person's style of response, as well as his product; for example, the examiner notes the sequence of movements, timing, comments and the like as the examinee draws a person or carries out an original dramatization. All of these techniques provide data that can serve as the basis for hypotheses about the individual that must be checked with the data obtained from other sources. Courses in the use of projective techniques require prerequisite training in clinical psychology; formal instruction must be supplemented by supervised experience in the administration, scoring, and interpretation of projective tests.

SELECTED REFERENCES

- ALLEN, R. M., *Personality Assessment Procedures: Psychometric, Projective, and Other Approaches*. New York: Harper & Row, Publishers, Inc., 1958.
- CAMPBELL, DONALD T., "A Typology of Tests, Projective and Otherwise," *Journal of Consulting Psychology*, vol. 21 (June 1957), pp. 207-210.
- CRONBACH, LEE J., "Response Sets and Test Validity," *Educational and Psychological Measurement*, vol. 6 (Winter 1946), pp. 475-494.
- , "Further Evidence on Response Sets and Test Design," *Educational and Psychological Measurement*, vol. 10 (Spring 1950), pp. 3-31.
- DIAMOND, SOLOMON, "The Factorial Approach," *Personality and Temperament*. New York: Harper & Row, Publishers, Inc., 1957, pp. 151-183.
- EDWARDS, ALLEN L., *The Social Desirability Variable in Personality Research*. New York: Holt, Rinehart and Winston, Inc., 1957.
- FRICKE, BENNO G., "Subtle and Obvious Test Items and Response Set," *Journal of Consulting Psychology*, vol. 21 (June 1957), pp. 250-252.
- HANLEY, CHARLES, "Social Desirability and Response Bias in the MMPI," *Journal of Consulting Psychology*, vol. 25 (February 1961), pp. 13-20.
- HENRY, WILLIAM E., "Projective Techniques," in Paul Mussen, ed., *Handbook of Research Methods in Child Development*. New York: John Wiley and Sons, 1960.
- JACKSON, DOUGLAS N., AND SAMUEL MESSICK, "Content and Style in Personality Assessment," *Psychological Bulletin*, vol. 55 (July 1958), pp. 243-252.

- LINDZEY, GARDNER, "On the Classification of Projective Techniques," *Psychological Bulletin*, vol. 56 (March 1959), pp. 158-168.
- MASLING, JOSEPH, "The Influence of Situational and Interpersonal Variables in Projective Testing," *Psychological Bulletin*, vol. 57 (January 1960), pp. 67-85.
- MESSICK, SAMUEL, *Measurement in Personality and Cognition*. New York: John Wiley and Sons, Inc., 1962.
- SUPER, DONALD E., AND JOHN O. CRITES, *Appraising Vocational Fitness*. New York: Harper & Row, Publishers, Inc., 1962, Chapter 19.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. Describe and evaluate one of the leading personality inventories used at the high school level. Study both the inventory and the manual, and consult the reviews in Buros' *Mental Measurements Yearbooks*.
2. Describe and evaluate a problems inventory for the high school level. Study the inventory and the manual, and consult the reviews in the Buros' *Mental Measurements Yearbooks*.
3. Compare the personality inventory and the problems inventory (mentioned in problems 1 and 2 above) with respect to method of construction, criteria used in validation, and purposes that they are designed to serve.
4. Select three tests of personality for children of the same age range and compare them on the basis of content. Describe what each test purports to measure.
5. Under what circumstances can transparent personality inventories be used to advantage? In which types of situations should their use be avoided?
6. Why should most projective techniques be used only as part of a case study by a qualified psychologist?
7. What cautions should be observed in the interpretation of children's art work?

PART THREE

The Improvement
of Instruction

Development, Try-out, and Revision of Teacher-Made Tests

As teachers have achieved a more significant role in planning the educational experiences for their classes, they have also become responsible for appraising the extent to which students are progressing toward the goals of the educational program. Even in those schools where standardized achievement tests are regularly administered, they are usually given only once a year; and they measure only a fraction of the educational outcomes. It is the teacher who is responsible for measuring student achievement day by day and week by week. He must develop his own tests for measuring student progress toward the immediate objectives of instruction.

IMPORTANCE OF TEACHER-MADE TESTS

It is through his own tests that the teacher communicates to students information concerning the knowledges and the intellectual skills that he considers most important. Tests provide students with tangible indications of the outcomes expected from a course, even to a greater degree than do the textbook or syllabus.

Teacher-made tests and other types of teacher evaluation constitute the basis for grading students and reporting to parents. It is largely teacher-made tests that provide students with confirmation or "feedback" concerning the effectiveness of their efforts to learn. The knowledge that a test is to be given provides most students with strong stimulation to study; the types of test questions used in previous tests direct their efforts to learning activities that they believe to be most helpful in improving their test performance. Teacher-made tests have great potentialities for enriching or limiting the students' self-directed study. If teacher-made tests faithfully

represent the major objectives of instruction, special studying or reviewing for tests will reenforce other aspects of teaching.

In the process of planning teacher-made tests and devising items for them, the teacher faces up to such questions as: (1) What kinds of student behavior would be evidence of progress toward each objective? (2) What type of test situation and what specific items would elicit from students this type of behavior? (3) How should such behavior be rated or scored? (4) Is this objective realistic? If students have not shown progress, do I know how to reorient or modify instruction so as to obtain better results? When teachers try to devise test items that will help them in judging student progress toward a major educational goal, they begin to see more clearly what the goal really means and how difficult it is to determine whether students are really making progress toward ultimate objectives.

CHARACTERISTICS OF A GOOD TEACHER-MADE TEST

The characteristics of a satisfactory measuring instrument, as outlined in Part One, are just as applicable to teacher-made tests as to standardized tests. Specific application of these criteria, however, depends on the purpose for which a test is to be used. Some teacher-made tests are used to measure the relative status of students in some aspect of achievement (the scores serving as a basis for assigning marks or otherwise ranking students). Others are designed chiefly to serve certain instructional purposes—identifying facts and processes that require reteaching, helping students to recognize gaps in their own achievement, and the like.

Tests Designed to Measure Relative Status

If a teacher is developing a test to aid in ranking students with respect to achievement in geography or some other area, the test should:

1. *be based on a representative sampling of the content studied.* The percentage of items on each topic should correspond approximately with the proportional emphasis given that topic in the course. In a geography test, for example, the teacher should consider the emphasis given in the course to each of the major geographic regions, as well as the relative stress placed on such topics as products, important cities, weather, soil conditions, trade routes, and the like.
2. *be based on a representative sampling of the abilities or skills emphasized in the course.* Memorization of facts may be the only skill involved in a test unless the teacher makes a special effort to include questions that require students to make comparisons, apply principles, read maps, and the like.

3. *contain a sufficient number of questions so that the test will have adequate reliability.* Although no exact rule can be given regarding number of items, a test used as a basis for grading should probably have a reliability coefficient of .70 or better.¹ Of course, a teacher can achieve the desired reliability by combining scores on a series of short tests (that is, recording and combining the raw scores, rather than assigning an A, B, or C grade to the student's scores on each short test).
4. *include items covering a wide range of difficulty.* The test items should range from easy to difficult, with a large number of items geared to the middle group and with several items difficult enough to challenge the best student. (Several easy items are also needed, but they will be all too prevalent without the teacher's trying to design them!)

Instructional Tests

If the teacher is using his test for group diagnosis, student self-evaluation, or other instructional purposes, and will not use the scores as a basis for ranking students, he need not be so concerned with the criteria listed above. For example, he may wish to include a disproportionate number of questions on a specific topic or geographic region because of its complexity or because he thinks that reteaching of certain aspects may be needed. This disproportionate emphasis would constitute a violation of criterion 1 and would distort the test scores as a basis for grading. The teacher may wish to emphasize certain skills to the exclusion of others (a violation of criterion 2). He may develop a short test that would be unreliable as a basis for judging the achievement of individual students on a topic but which would serve to show what facts or concepts had been learned by the group as a whole. Or he may use a short test to help students clarify the more important learnings from a motion picture or excursion (see criterion 3). In a test on minimum essentials, the teacher would not be much concerned with range of difficulty of items; in fact, he might wish to include a large number of items that he hoped almost everyone had learned in order to assure himself that all students had mastered these basic concepts and to identify the few who had not (see criterion 4). That is, an instructional test that was not designed to rank students on the basis of their total scores might violate one or more of the criteria listed above and still be a valuable classroom test.

Superficially, it might seem that the instructional test does not need to meet any standards. Every evaluation device, however, should meet the basic criteria of validity, reliability, and usability in terms of the purpose

¹ Although higher reliability is desirable, it is seldom achieved in teacher-made tests. Increased reliability can be achieved by combining scores from several tests given during the semester or year. Table 3.5 and the quick method of estimating *SD* (Table 2.1) can be used to estimate reliability coefficients quickly.

for which it is being used. For example, any test designed to help diagnose certain types of errors in arithmetic should contain several problems of each type. Thus, student scores can be related to types of weaknesses, rather than representing chance errors. All tests should be concerned with major outcomes rather than trivialities. All tests should be so skillfully constructed and scored that extraneous factors do not invalidate scores. Questions should not be ambiguous; the intent of each question should be clear to all students who are prepared for the test. Catch questions, "textbook language," and stereotyped verbalizations should be avoided.

Brownell has developed several criteria for judging the worth of classroom tests in relation to the instructional process:

Does the test elicit from the pupils the desired types of mental processes?

A test is good to the extent that it calls forth the mental processes which should be measured. . . . It is not enough to be able to recall facts; the facts learned must function . . . we must deliberately set out first to teach the desired mental processes, and then measure the degree to which instruction has developed these processes.

Does the test encourage the development of desirable study habits?

The type of test given determines the "set" adopted by the pupil in his study. If this set is made habitual by reason of the teacher's more or less exclusive use of a particular type of test, the pupil builds his study habits accordingly and neglects other valuable methods of attack upon subject matter.

Does the test lead to improved instructional practice?

. . . just as the student adapts his study procedures to the type of test which has been announced, so the teacher adapts her instruction to fit the type of test she intends to give.

Does the test foster wholesome relationships between teacher and pupils?

No one who has observed children under test conditions can doubt the possible effects of such situations upon their mental health . . . tests, which are pleasurable experiences to young children, become events to be dreaded and avoided by older children. . . . If a test in any way impairs healthful, wholesome relationships and growth toward integrated personality, that test is bad.²

² William A. Brownell, "Some Neglected Criteria for Evaluating Classroom Tests," *Appraising the Elementary School Program*, 16th Yearbook, Department of Elementary School Principals (Washington, D.C.: Copyright 1937, National Education Association) pp. 485-492.

In these criteria, Brownell emphasizes the close relationship between evaluation and instruction. He recognizes that a teacher's chief concern, as he develops his own tests, should be that the evaluation advance the quality of instruction. If the teacher's tests stimulate his students to study relationships and apply principles, if they encourage the development of desirable study habits, and if they are accepted by students as both fair and helpful, testing can contribute immeasurably to the effectiveness of instruction. The teacher's discussion of test results when they are returned may determine whether testing facilitates or impedes the achievement of educational goals.³

All tests should avoid overemphasis on isolated facts, as opposed to ideas and concepts that have more general application. Hawkes and others suggest emphasizing questions involving "why, wherefore, how, with what results, of what significance, explain, interpret, and compare," as opposed to "who, what, when, describe, and name."⁴

Information is important; but facts, terms, and rules are best learned when they are meaningfully interrelated to important concepts and principles that have meaning to the student, which he can state in his own words and apply to new situations. Certainly in daily quizzes or weekly tests, the examination of students' recall of specific facts and terms is appropriate. Even here, however, items that test the students' comprehension, rather than just his memory, encourage a type of studying that leads to longer retention and greater probability of transfer.

When major units of work are completed, the examination should certainly require students to demonstrate their ability to apply concepts and principles in test situations that are somewhat unfamiliar, that is, that do not exactly parallel the textbook problems.

PLANNING TESTS FOR GREATER CONTENT VALIDITY

Dressel has devoted many years to leadership in the field of relating evaluation and instruction, showing teachers how carefully formulated objectives can serve as bases for developing both learning experiences and evaluation procedures. The following listing of parallel elements in these two processes clarifies how closely they should be interrelated:

³ A helpful article on teacher discussion of achievement test results is Roger T. Lennon, "Testing: Bond or Barrier between Pupil and Teacher," *Test Service Bulletin* No. 82 (New York: Harcourt, Brace & World, Inc., n.d.).

⁴ Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann, eds., *The Construction and Use of Achievement Examinations: A Manual for Secondary School Teachers*. (Boston: Houghton Mifflin Company, 1936), p. 111.

INSTRUCTION

1. Instruction is effective as it leads to desired changes in students.
2. New behavior patterns are best learned by students when the inadequacy of present behavior is understood and the significance of the new behavior patterns thereby made clear.
3. New behavior patterns can be more efficiently developed by teachers who know the existing behavior patterns of individual students and the reasons for them.
4. Learning is encouraged by problems and activities that require thought and/or action by each individual student.
5. Activities that provide the basis for the teaching and learning of specified behavior are also the most suitable activities for evoking and evaluating the adequacy of that behavior.

EVALUATION

1. Evaluation is effective as it provides evidence of the extent of the changes in students.
2. Evaluation is most conducive to learning when it provides for and encourages self-evaluation.
3. Evaluation is conducive to good instruction when it reveals major types of inadequate behavior and the contributory causes.
4. Evaluation is most significant in learning when it permits and encourages the exercise of individual initiative.
5. Activities or exercises developed for the purposes of evaluating specified behavior are also useful for the teaching and learning of that behavior.⁵

Ideally a teacher who is designing a comprehensive examination should define the universe to be sampled in much the same way as the author of a standardized test. That is, he should decide the proportional emphasis that should be given to different *content* areas so as to represent his instruction fairly; and he should also decide the *types of abilities* to be sampled by the test items (for example, recognition or recall of learned material, comprehension as shown by some type of interpretation, use of learnings in new situations, and the like). That is, his plan for the test should consider the relative emphasis to be given both to *content* areas and to *processes* or *cognitive abilities* (specific ways of responding to, or dealing with, the course content).

⁵ P. L. Dressel, "Evaluation as Instruction," *Proceedings of the 1953 Invitational Conference on Testing Problems* (Princeton, N. J.: Educational Testing Service, 1954).

An Illustrative Table of Specifications

If the teacher accumulates test items without a plan, they will unduly represent informational learnings, especially knowledge of specific facts. Moreover, teachers are likely to overemphasize certain areas of content in which items are easily constructed. In order to improve the test's representativeness, or its content validity, one should first develop a blueprint for the test.

Table 10.1
Specifications for a Final Examination in Natural Science—Term 3

Objectives ^a	Knowl- edge	Comprehension (Translation, Interpretation, Extrapolation)	Appli- cation	Analysis	Total
COURSE CONTENT					
I. The number concept	6	4			10
II. Fundamental concepts of arithmetic	4	4	2		10
III. Quantitative descriptions	2	3	3	2	10
IV. The gas laws	3	2	2	3	10
V. "The spring of the air"	1	2	3	4	10
VI. The kinetic theory of matter	4		4	2	10
VII. The theory of the atom	1	2	4	3	10
VIII. Electricity and combustion	1		6	3	10
IX. Static electricity and magnets	2	2	3	3	10
X. Electricity and the nature of matter	1	1	3	5	10
Total	25	20	30	25	100

^a Based on Benjamin S. Bloom, ed., *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain* (New York: David McKay Company, Inc., 1956).

Source: Used with the permission of the publisher from Paul L. Dressel and Associates *Evaluation in Higher Education* (Boston: Houghton Mifflin Company, 1961), p. 119.

In Chapter 4, in our discussion of content validity, we presented an illustrative test blueprint. In Table 10.1, we present a blueprint, or table of specifications, for a science examination, in which the items are classified in terms of *both* content and objectives. In the left-hand column are listed the ten content areas to be represented by test items. Across the top are listed four types of educational goals that constitute the first four major categories of the taxonomy of educational objectives. The teacher designing this final examination has decided that 25 percent of all test items should test progress toward knowledge goals; 20 percent, comprehension; 30 percent, application; and 25 percent, analysis. No attempt was made to attain this proportion of items in each content category, but rather in the test as a whole.

Content-Orientation vs. Goal-Orientation

This two-way table of specifications gives adequate consideration to both the content and cognitive abilities of the course. It is important that test coverage be adequate from both points of view.

The history of work on educational objectives has been an interesting one, and one that has had great implications for evaluation. In the first wave of enthusiasm for behaviorist psychology, just following World War I, objectives were stated in highly specific and utilitarian terms (the ability to multiply specific number combinations, to spell specific words, to recall the symbols for specific elements, and the like). Within a decade, this approach to educational objectives fell into disuse, partially because of the tremendous volume of specifics and partially because new studies in transfer of training had revealed that the learning of general principles and the development of generalized behaviors resulted in greater economy of learning, and greater interest and retention, than the learning of isolated specifics.

When the importance of learning for transfer was recognized, lists of highly generalized objectives were then developed by national associations in the various subject fields. The following list for social studies is an example:

1. Acquisition of important information
2. Familiarity with technical vocabulary
3. Familiarity with dependable sources of information on current social issues
4. Immunity to malicious propaganda
5. Facility in interpreting social science data
6. Facility in applying significant facts and principles to social problems of daily life

7. Skill in investigating social science problems
8. Interest in reading about social problems and in discussing them
9. Sensitivity to current social problems
10. Interest in human welfare
11. The habit of working cooperatively with others
12. The habit of collecting and considering appropriate evidence before making important social decisions⁶

Since that time some educators have felt that teachers committed to major objectives need only the freedom to plan instructional activities, keeping these major objectives in mind; while others have realized that a tremendous amount of study and evaluated experience is needed to discover activities and materials that will help students achieve intermediate objectives related to these ultimate objectives.

As instructional programs have become more diversified, published tests with a subject-matter orientation have come to fit fewer and fewer courses. Test publishers have shifted more to the measurement of developed intellectual abilities and/or student progress toward the ultimate goals of education.

It is essential, however, that students be fairly graded on the knowledge objectives of a specific course; that the adequacy of their learning of specifics be measured so that correct learnings can be reinforced and gaps in achievement be identified. Teacher-made tests (developed by one teacher or by several teachers working cooperatively) must bear much of the burden of measuring student learnings in knowledge. The specific facts taught to clarify basic principles can vary to some extent from one locality to another. It is appropriate, therefore, that teacher-made tests, rather than standardized tests, bear much of the burden of testing students for knowledge of specifics.

The controversy between content-oriented and goal-oriented teachers regarding the importance of teaching and testing knowledges might approach a constructive solution if both groups would recognize the fallacies of either extreme point of view. Content-oriented teachers can become so preoccupied with the importance of students' learning specific facts and terms that their students fail to gain a sense of direction and a level of understanding, which comes from seeing the interrelatedness of concepts and principles and applying them to unfamiliar problems. On the other hand, goal-oriented teachers who focus their attention on long-range objectives may fail to give adequate attention to the specific learnings that help the learner along the road to the ultimate goal. As Dressel says, they "some-

⁶ *The Social Studies Curriculum*, Yearbook, Department of Superintendence (Washington, D.C.: National Education Association, 1936), pp. 320-340.

Table 10.2
Relative Advantages of Essay and Other Supply-type Test Items^a

ADVANTAGES	DISADVANTAGES
<ol style="list-style-type: none"> 1. Easily prepared <ol style="list-style-type: none"> a. Fewer questions to prepare b. Need not be mimeographed 2. Largely eliminates guessing. 3. Stimulates use of superior study methods in preparation (as compared with study methods used in preparing for objective tests). 4.* Represents the most direct method of testing many outcomes. For some objectives, test exercises can closely approximate criterion behavior. 5.* Provides more adequate basis for making inferences about student's <i>level</i> of competency, for example, his ability to define words, write clear explanations of procedures, or complete a geometric proof. 6.* May give student opportunity to demonstrate his ability to: <ol style="list-style-type: none"> a. Choose most pertinent and important learnings b. Organize his knowledge c. Express opinions and attitudes d. Show initiative and originality 7.* May be useful for diagnosing incorrect interpretations and partially understood concepts. 	<ol style="list-style-type: none"> 1. May have relatively low reliability, owing to <ol style="list-style-type: none"> a. Limited sampling of learnings b. Subjectivity of scoring 2. May have relatively low validity, owing to <ol style="list-style-type: none"> a. Limited sampling of learnings b. Low reliability 3. Requires excessive time of students in writing 4.* Tends to be quickly and carelessly constructed, with the result that: <ol style="list-style-type: none"> a. Questions are ambiguously stated b. Questions are so general that a student can bluff or "talk around" the subject c. Questions are of unequal difficulty, with no provision for weighting them unevenly in the scoring process d. Selection of questions is not representative of major learnings 5.* Tends to be graded without an adequate scoring key, so that students' marks are affected by <ol style="list-style-type: none"> a. The "halo effect" (good or bad) of a student's previous level of performance or success on a single question of the test b. Legibility of handwriting c. Errors of spelling and grammar d. Effectiveness of written expression

^a Variable factors, which are markedly affected by the teacher's skill and care in test construction and scoring.

Table 10.3
**Relative Advantages and Disadvantages of Objective
 or Selection-type Items**

ADVANTAGES	DISADVANTAGES
<ol style="list-style-type: none"> 1. Makes possible extensive sampling of learnings in a relatively short testing time. 2. Tends to have high reliability as the result of <ol style="list-style-type: none"> a. Objectivity of scoring b. Extensiveness of sampling 3. Is scored objectively, with the result that: <ol style="list-style-type: none"> a. Scoring time is reduced b. The teacher is freed from suspicion of partiality c. Students' scores are not affected by such extraneous factors as ability to write rapidly and the like d. Students may do self-scoring e. Scoring may be delegated to clerical workers or readers f. Statistical analysis of student performance (on the test and on specific items) is facilitated g. At the upper grade levels, tests can be scored by machine and summaries made of student success on each test item 4. Focuses student attention on specific facts or abilities being tested, permitting no evasion. 5.* May be valuable for such instructional purposes as: <ol style="list-style-type: none"> a. Pretesting b. Diagnostic testing c. Individualized self-testing d. Testing for application of principles to new situations 	<ol style="list-style-type: none"> 1. Requires time and skill for adequate preparation. (<i>Note:</i> Good test questions, however, can be reused. Moreover, time spent by the teacher in test preparation may result in increased awareness of the goals of instruction, major concepts to be developed, facts related to such concepts, and the like). 2. Is of limited value in some subjects. 3.* May stimulate superficial learning of many details because of <ol style="list-style-type: none"> a. Emphasis placed on <i>recognition</i> of correct answer rather than on remembering b. Failure to require the student to organize significant facts and ideas and to reason about them 4.* May include ambiguous or misleading questions. 5.* May result in unduly high scores for intelligent, test-wise students, who have <i>not</i> studied, because of: <ol style="list-style-type: none"> a. Transparent clues in grammar, word form, or phrasing b. Insufficient care in the devising of incorrect responses, allowing correct answers to be chosen more by the avoidance of obviously poor choices than by recognition of the correctness of the right response

* Variable factors, which are markedly affected by the teacher's skill and care in test construction and scoring.

times forget that a compass does not relieve the traveler from the necessity of choosing routes and means of transportation."⁷

The ultimate objectives of a course are usually stated in such generalized terms that a teacher must think through the specific learnings that contribute to their achievement before he has an adequate basis for planning his tests or other evaluation instruments.

THE ADVANTAGES AND DISADVANTAGES OF ESSAY AND OBJECTIVE TESTS

The development of objective tests in the period following World War II was stimulated by research studies that revealed the very low reliability of student scores on the traditional essay test. For a period of time following these discoveries, the essay test was almost universally in disrepute, and the new objective test was almost as consistently praised. Today educators recognize the strengths and limitations of each approach. Moreover, there is a growing recognition that many of the criticisms of both approaches are not necessarily inherent but grow out of ineffectiveness in their application.

The advantages and disadvantages of essay and objective tests are summarized in Tables 10.2 and 10.3. Those advantages and disadvantages that *tend to characterize* one approach or the other are listed first, followed by other factors that are markedly affected by the teacher's skill and care in test construction and scoring. Many of the claims made for the essay examination, for example, are not realized in practice; whereas many of the criticisms of the essay examination can be minimized by care in construction and scoring. Similarly, the objective examination has potentialities that may or may not be realized, depending upon the skill used in test construction.

Ideally, a teacher uses a combination of the two approaches and is constantly trying to improve his effectiveness in each. In fact, teachers are increasingly combining objective and essay questions in a single test in order to obtain both the advantages of the former in terms of more extensive sampling, higher reliability, and objective scoring and the advantages of the latter in stimulating superior study methods and giving the student opportunity to organize his knowledge and express his own opinions and attitudes.

⁷ Paul L. Dressel, "Measurement and Evaluation of Instructional Objectives," *17th Yearbook*, National Council on Measurements Used in Education. (New York: The Council, 1961), p. 4.

THE CONSTRUCTION OF TEST ITEMS

If he is to develop good tests, the teacher must develop skill in the construction of the major types of test items. Research studies have revealed no order of merit in the various types of test items. Proficiency in writing all types will enable the teacher to select the best type of test question for each of the various outcomes to be tested.

Supply-Type Items: Essay Questions

Teachers like essay questions because of their ease of preparation, but dislike them because of the time required for scoring and the difficulties of explaining grading to students. The teacher who is willing to spend time in the careful formulation of essay questions so that they focus clearly on the basic principle or concept to be tested will achieve some of the special values claimed for the essay examination and will also save time in scoring the test and interpreting the results.

Remmers and Gage offer the following suggestions for improving essay examinations:

1. Use essay questions to evaluate achievement of only those instructional objectives not as well or better tested by the short-answer forms.

2. Phrase the questions so as to require as precisely as possible the specific mental processes operating on specific subject matter. . . .

3. . . . phrase the questions so as to give as many hints concerning the organization of the pupil's answers as are not inconsistent with the instructional objective at which the questions are aimed. . . . the more specific the essay question becomes, the more similar it becomes to short-answer test items. Carried to an extreme, this technique would rob the essay question of its unique value in testing the pupil's ability to organize and express his answers. . . . We can attempt to elicit as much organizational effort from the pupils as possible while giving them a common set of reference points so that their answers will be comparable.

4. Permit no choice among questions. Only by requiring all pupils to answer all questions can their achievement be compared. . . . The teacher who permits pupils to choose among optional questions can never know whether all of them have taken a test of equal difficulty. . . .

5. Balance the questions in difficulty so that the pupil can actually write adequate answers to all of them within the allotted time if he possesses the required achievement.⁸

⁸ H. H. Remmers and N. L. Gage, *Educational Measurement and Evaluation* (New York: Harper & Row, Publishers, Inc., 1955), pp. 183-184.

Since essay tests require considerable time to score, we should construct questions that will not require students to provide a great deal of background information which almost all of them know. If an essay test item is to justify the scoring time, it should be so constructed that students devote most of their response time to those aspects of the question that will differentiate most effectively among them with respect to their achievement of some significant outcome of instruction. The students in a measurement class, for example, may all know the headings of the taxonomy if this information has been stressed by the instructor. A question that would require them to use a mimeographed list of headings of the taxonomy to classify test items and justify their choices would differentiate more effectively among them, and on a more significant basis, than a question merely requiring memory of its major headings and subheadings.

Before grading students' responses, the teacher should actually take the essay examination himself, listing the points that he expects students to make in response to each question. Other acceptable points made by students can be added to the key as scoring progresses. If the question cannot be analyzed into parts but must be rated as a whole, a sorting process is recommended. In using this method the teacher might sort students' responses into piles representing the five letter grades and intermediate degrees of merit, for example, "A," "between B and A," "B," and the like. The teacher should later reread those papers that are grouped as "between A and B," and all similar classifications, as a basis for reassigning them to one grade or the other.

The teacher should make every effort to avoid the "halo effect" in scoring—that is, the effect on a test grade of the teacher's attitude toward a student (because of the student's behavior or his general level of past performance). Such procedures as the following are helpful:

1. Keeping the identity of students secret during grading.
2. Correcting question 1 on *all* papers; then question 2 on *all* papers, and the like.
3. Occasionally reshuffling the papers so that a student's paper may not be graded unduly low or high because of its position after an unusually good or poor paper.

Since one of the purposes of essay questions is to encourage originality, the teacher should accept and encourage original answers that are based on facts and show good thinking.

Supply-Type Items: Recall or Completion Questions

The most widely used recall items are of two types: a direct question that can be answered by a single word or a short phrase, such as "Who

invented the steamboat?" and a simple sentence presented in incomplete form, such as "The steamboat was invented by _____."

The recall question shares with the essay question the advantages of ease of construction and the fact that the student is asked to recall information, rather than merely to recognize the correct response. If the teacher words his questions so as to avoid ambiguity and constructs a scoring key that includes all acceptable answers, the scoring can be highly objective. Because of its very nature, however, the use of the recall question tends to be limited to the testing of descriptive information and associations of the who, what, when, and where type.

The following suggestions may assist the teacher in improving the quality of his recall items:

1. In general, use recall items only when the correct response is a single word or brief phrase.
2. In recall questions of the completion type, it is best to omit only one *key* word or phrase; the omitted word or phrase should preferably be at the end of the sentence.
3. Avoid indefinite statements. Be sure that the kind of response wanted is clearly indicated; for example, "The steamboat was invented in the year——" rather than "The steamboat was invented in——." (The locality instead of the date might conceivably be given in response to the second question.)
4. Make minimal use of stereotyped phrases or other "textbook language"; avoid placing a premium on the student's recalling a unique word or phrase when other responses would indicate understanding of the concept.
5. Avoid having the grammatical structure of the question or statement give a clue to the response. Use of the article "a" or "an," for example, will provide students with a clue.
6. Do not give clues to the answer by varying the number or length of the blanks. For example, use "Florida was explored by _____," rather than "Florida was explored by—— ———."

When the teacher wants to check on the student's recall of specific terms and facts that are *basic to further work*, the recognition-type of question may be quite inadequate. Many teachers either use recognition items as admittedly inadequate substitutes or spend hours scoring recall questions. By using the ingenious procedure illustrated in Figure 10.1, the teacher can require knowledge at the recall level for significant terms and facts. Such answer sheets can be easily and objectively scored by a lay-over scoring stencil; or standard answer sheets providing room for as many as 12 choices can be scored by machine (each student being supplied with a "marker" indicating the way in which the letters of the alphabet have been allocated to the answer spaces).

Directions: After you have answered all completion questions in the usual manner, carefully indicate your answers on this sheet by blackening in for each item the answer space corresponding to the *third* letter of your response. If the name of a person has been requested, use his last name only; ignore an apostrophe, hyphen, or space; that is, the third letter in O'Brien is *r*.

EXAMPLES

A. For what term is the formula

$$\sqrt{\frac{\sum x^2}{N}} \text{ used?}$$

A.  AB CD EF GH IJK LM N OPQ R S TU VW XYZ

Since the answer is standard deviation, the first answer space is marked.

B. What term is used for the expression under the radical sign?

B.  AB CD EF GH IJK LM N OPQ R S TU VW XYZ

Since the answer is variance, the 9th answer space is marked.

Do *not* guess; a penalty for guessing is used in the scoring process. Remember that you are indicating the *third* letter.

1.  AB CD EF GH IJK LM N OPQ R S TU VW XYZ

11.  AB CD EF GH IJK LM N OPQ R S TU VW XYZ

10.  AB CD EF GH IJK LM N OPQ R S TU VW XYZ

20.  AB CD EF GH IJK LM N OPQ R S TU VW XYZ

Fig. 10.1 Answer Sheet Used To Facilitate Scoring of Completion Questions.

This approach was devised by C. F. Willey, "Fully Objective Scoring of the Completion Test," A paper presented at the 1963 Convention of the American Educational Research Association. Reported, in part, in C. F. Willey, "Objective Scoring of the Completion Test," *Psychological Reports*, vol. 10 (April 1962), pp. 501-502. The first letter is not used since the student might have a hazy recollection of a term and be able to recall the first letter. The second letter is so frequently a vowel that the third letter seems best.

Selection-type Items

Before we discuss true-false, multiple-choice, and matching items, it is well to list some general suggestions for item-writing that apply to all selection-type items.

1. All test items should be clearly and simply worded and should be grammatically correct.
2. Stereotyped expressions and textbook language should be avoided. Statements should not be "lifted" from the textbook.
3. Questions should be edited to reduce ambiguity of wording; there should be only one way in which a statement or term could be interpreted by students.
4. One should avoid providing clues to the right answer, for example, having the right answer longer or more cautiously phrased or using such "specific determiners" as are listed in the sections on true-false and other recognition-type questions.

TRUE-FALSE ITEMS Of the various types of questions found in teacher-made tests, the true-false item is undoubtedly the most widely used and the most severely criticized. True-false items are popular with teachers because they seem relatively easy to construct. A large number of true-false items can be typed on a single page and can be answered by students in a few minutes of class time. Scoring is rapid and easy. Unless they are carefully constructed, however, true-false questions are likely to be either obvious or ambiguous. Moreover, since a student has a 50-50 chance of answering any question correctly by guessing, true-false questions are of little help in diagnosis. Many gaps in the student's knowledge may be concealed by successful guesses.

If they are carefully constructed, however, true-false questions have a definite and important place in teacher-made tests. They make it possible to test a relatively large sampling of learnings per unit of student time. Moreover, true-false items are very well adapted to testing (1) understanding of principles or generalizations; (2) persistence of popular misconceptions; and (3) situations in which there are only two logical responses (north, south; right, left; colder, warmer; larger, smaller; and the like). Examples of each of these three types are given below:

Type 1. The best way to keep farm soil in good condition is to plant the same crops each year.

Type 2. Swallowing grape or watermelon seeds causes appendicitis.

Type 3. Most of the people of the world live in the Northern Hemisphere.

True-false questions can be used to advantage in instructional tests, especially if the administration of such test questions is followed by class discussion. Since many generalizations are neither wholly true nor wholly false, the symbol *S* can be added for "sometimes true and sometimes false."

T F S 1. Two right isosceles triangles are congruent if a leg of one equals a leg of the other.

T F S 2. Two isosceles triangles are similar if any angle of one equals the corresponding angle of the other.

T F S 3. An equilateral quadrilateral is a square.⁹

⁹ Hawkes, Lindquist, and Mann, *op. cit.*, pp. 372-373.

In science, industrial arts, and other subjects, true-false questions can be used to test the student's ability to apply principles to new situations. For example, the following questions are based on a diagram showing electrical circuits and switches:

- Yes No 1. If switch *D* is closed, will it cause a short circuit?
 Yes No 2. If switch *S* is open, will there be a flow of current?
 Yes No 3. If switches *S* and *A* were closed, would this create a short circuit?
 Yes No 4. Can light *B* be controlled by switch *D*?
 Yes No 5. Does the current in this circuit flow from *X* to *Y*?¹⁰

In instructional tests, many teachers prefer an adaptation of the true-false question that requires the student to indicate why a statement is false or to revise the statement so as to make it true.

Directions. Some of the following statements are true and some are false. If the statement is true, encircle the "T" at the left and do no more. If the statement is false, encircle the "F" and do two more things:

1. In blank "A" insert the word that makes the statement false.
2. In blank "B" insert the word that would make it true.

DO NOT USE WORDS THAT ARE UNDERLINED. The first item is answered as an example.

- (X) T (F) Large city newspapers are printed on cylinder presses.

A. _____ (cylinder) B. _____ (rotary)

- (1) T F The optical center of a page lies just below the true center.

A. _____ B. _____

- (2) T F Fir plywood is sold by the board foot.¹¹

A. _____ B. _____

In order to retain the values inherent in this type of question and still have the advantages of objective scoring, the following variation of the true-false question can be used:

Directions. Some of the following statements are true and some are false. If the statement is true, encircle the "T" preceding the item and do no more. If the statement is false, encircle the "F" and do two things:

1. Underline the word that makes the statement false.
2. From the list just below, select the word that would make the item true and place the letter preceding that word in the blank space before the item.

¹⁰ From William J. Micheels and M. Ray Karnes, *Measuring Educational Achievement*. Copyright 1950. McGraw-Hill Book Company, Inc., p. 208. Used by permission.

¹¹ From William J. Micheels and M. Ray Karnes, *Measuring Educational Achievement*. Copyright 1950. McGraw-Hill Book Company, Inc., p. 203. Used by permission.

The first item is answered as an example.

A. 10	G. attract	M. parallel
B. 15	H. cell	N. repel
C. 25	I. copper	O. series
D. alternating	J. current	P. steel
E. aluminum	K. direct	Q. voltage
F. ampere	L. ohm	R. watt

(X) T ☒ (L) The volt is the unit of resistance.

(1) T F _____ If a DC circuit has a pressure of 20 volts and a current of 5 amperes, the resistance would be 4 ohms.

(2) T F _____ The elements of a telegraph circuit are connected in parallel.¹²

The following suggestions may assist the teacher in constructing better true-false items:

1. The questions should be related to significant facts or generalizations. The best true-false statements require the student to understand a significant fact or generalization presented in a new way.
2. The crucial element in the statement should be readily apparent to the student. Ordinarily it should be placed in the main clause and near the end of the statement. Underlining the crucial word or words may be desirable.
3. Avoid "lifting" true statements directly from the textbook or developing false statements by the mere insertion of the word "not" into such a lifted statement. Not only does such a procedure encourage rote learning, but some textbook statements are ambiguous when removed from their context.
4. Avoid the use of specific determiners, words, or phrases that are usually associated with either a true or a false statement. Such words as "all," "none," "always," "never," and the like are usually associated with false statements; whereas statements containing "some," "generally," "may," "should," and the like are usually true.
5. Avoid making true statements consistently longer than false ones.
6. Have a somewhat larger number of false than true statements. This suggestion is made because the student who does not know the answer is more likely to guess "True" than "False."
7. Avoid statements that are partly true and partly false.
8. Speed up scoring by typing the symbols T and F in a column (preceding or following the questions) so that students can mark their choices, and a lay-over scoring stencil (with holes in the positions for correct responses) can be used. If no answer column has been typed on the test, or if the test questions are dictated, have students write the symbols + and 0. These are more easily distinguished in scoring than T and F or + and -; they are also less easily changed by students when self-scoring is used.

MULTIPLE-CHOICE ITEMS In a multiple-choice item, either an incomplete statement is followed by several possible completions; or a direct

¹² From William J. Micheels and M. Ray Karnes, *Measuring Educational Achievement*. Copyright 1950. McGraw-Hill Book Company, Inc., p. 208. Used by permission.

question is followed by several possible answers. Typically, a multiple-choice question is designed so that only one answer is correct. However, variations have been developed in which (1) the student selects the *best answer* from a number of responses that vary in their acceptability; (2) the student selects one *incorrect* or otherwise inappropriate response from a group of three or more; or (3) the student checks *two or more correct responses* from a list of several alternatives.

Multiple-choice items can be designed to require reasoning and judgment as well as a knowledge of facts. Multiple-choice items that are well constructed tend to be much more valid and reliable than an equivalent number of true-false items. They are applicable to evaluating growth toward a wide variety of instructional goals. They are usually well liked by students. They can be easily and objectively scored.

On the debit side should be mentioned the fact that good multiple-choice items are difficult to construct. They require more student time in responding and more space on the page than do true-false or recall items. That is, the greater reliability of multiple-choice items is counterbalanced in part by the smaller number of items (and therefore the smaller sampling of learnings) that can be tested in the period of time required by a much larger number of true-false questions.

Because of their greater difficulty of construction, multiple-choice items should probably *not* be used when a simple recall item is adequate—that is, when there is clearly only one correct response and that response is a single word, number, or brief phrase; or when there are *only two* plausible responses (for example, right or left, North Pole or South Pole, safe or unsafe, and the like). In the latter case, a true-false item is usually effective.

The value of a multiple-choice item depends largely on the skill with which the incorrect choices, or distractors, are written. For some multiple-choice items, five plausible alternatives can be constructed; for others, the teacher may have only three plausible choices. In the latter case, the inclusion of two additional choices that are obviously false adds nothing to the measurement value of the question.

Although the number of available choices should not vary at random throughout a test, there is no reason why a teacher cannot use the multiple-choice technique for a group of questions where there are as few as three responses. For example, in a science test, a situation might be described and a series of conditions listed, for each of which the student would check one of *three* alternatives; “increases growth,” “decreases growth,” or “no change.”¹³

¹³ Williams and Ebel studied the effects on test reliability of omitting the least plausible alternatives from the items of an expertly constructed four-choice vocabulary test. Reducing the number of alternatives to three increased students' speed of working, while reducing them to two increased their speed considerably. For tests of

The following suggestions may be helpful in improving the quality of multiple-choice items.

1. As much of the item content as possible should be put in the stem of the item.¹⁴ If this is done, the informed student will have the answer in mind before he scans the options given. Moreover, space and student time are conserved because repetition of words in the various options is avoided.
2. The inexperienced item writer may find it advisable to use the direct question rather than incomplete sentence form, since the question form forces him to state the problem clearly and also reduces the risk of giving the student clues through grammatical inconsistencies. However, the more experienced item writer will prefer the incomplete sentence, because careful phrasing of the stem may reduce the length of the options.
3. Make all responses plausible. It should be necessary for the student to read and consider all choices presented. The alternative choices should deal with the same family of ideas—that is, should be reasonably homogeneous with respect to period of history, geographic area, or other basis of classification. For example, in the question

Which of the following men invented the telephone: (a) Edison; (b) Bell; (c) Marconi; (d) Morse?

all four choices are *inventors* in the field of *communications* in the *same period of history*; hence, the question is a better one than if the responses were less homogeneous.

4. The correct answer should not be consistently longer than the incorrect ones.
5. Avoid giving clues to the unprepared student through grammatical construction or other means. The incomplete sentence is especially likely to include grammatical clues—use of “a” or “an,” use of singular or plural subject or verb, and the like. All options must be grammatically consistent with the stem. The question

The explorer who claimed the Mississippi Valley for France was: (a) Pizarro; (b) LaSalle; (c) Cabot; (d) Hudson; (e) Smith.

is a one-choice, rather than a five-choice, question for a student who can identify French names.

6. Avoid the use of textbook language or stereotyped phrases.
7. The position of the correct answer should be randomized throughout the test.

Many teachers find that administering a group of recall or completion items often produces a number of plausible incorrect responses, which

equal working time, three-choice items gave a test of equal reliability, and two-choice items gave a test of slightly higher reliability. B. J. Williams and Robert L. Ebel, “The Effect of Varying the Number of Alternatives per Item on Multiple-Choice Vocabulary Test Items,” *14th Yearbook, National Council on Measurements Used in Education* (New York: The Council, 1958), pp. 63–65.

¹⁴ The direct question or the incomplete statement that poses the problem is called the “stem” of the test item.

they can use in developing a multiple-choice test for later use. Incorrect responses obtained in this way are often more effective than ones that the teacher could invent, since they represent genuine sources of confusion for the students. In mathematics, the false alternatives in multiple-choice questions can be selected so as to represent typical errors or misunderstandings by students.

MATCHING EXERCISES A matching exercise is a special type of multiple-choice question. The usual multiple-choice question has a single problem or stem, followed by two to five options. In a matching exercise, there are *several* problems or questions; the answer to each one is to be chosen from a single list of options. That is, the same list of alternative responses is used for several test items included in one matching exercise. Obviously, a matching exercise has the advantage of compactness because it requires less space on the page than multiple-choice questions based on the same content. Such exercises are rather easily constructed and are easily and objectively scored. In the following example, each of the options can be used more than once.

Directions: Listed below are several types of lubricants, followed by automotive units that require one of these lubricants. You are to match each unit with the correct type of lubricant and place the identifying letter in the blank space provided. Use each letter (A, B, and the like) as many times as is necessary. The first item is answered as an example.

- | | |
|--------------------------------------------------|-----------------------------------|
| <u>(C)</u> (X) transmission | A. engine oil |
| ___ 1. distributor | B. fibrous grease |
| ___ 2. striker plate | C. gear oil |
| ___ 3. universal joint | D. lubricant impregnated |
| ___ 4. dovetail | E. penetrating dripless lubricant |
| ___ 5. differential | F. pressure gun lubricant |
| ___ 6. door hinges | |
| ___ 7. generator | |
| ___ 8. front wheel bearings | |
| ___ 9. drag-link ends | |
| ___ 10. spring pins | |
| ___ 11. steering gear | |
| ___ 12. carburetor air cleaner | |
| ___ 13. spindle pin | |
| ___ 14. drive shaft center bearing ¹⁵ | |

Matching questions, however, have their special limitations and hazards. They are not well adapted to testing student knowledge in small units of subject matter; it is difficult to find items that are sufficiently homogeneous to require much discrimination on the part of students. Matching exercises are more likely than any other type of objective test item to include irrelevant clues to the correct response.

¹⁵ Adapted from Micheels and Karnes, *op. cit.*, p. 233.

The items in at least one of the two lists of a matching exercise should consist of single words or numbers or very brief phrases. Hence, matching exercises are well adapted to who, what, when, and where types of learnings and are usually considered ineffective in testing for understandings. An exception is the matching of items to indicate cause-effect relationships, as in the following examples:

COLUMN 1 (Statements of Effects)	COLUMN 2 (Statements of Causes)
(C) (X) Fishing in Norway () 1. Wandering life of the Eskimos () 2. Transportation by camel () 3. Houses of wood and bark with steep roofs of leaves () 4. People wearing wooden shoes when working	A. High, snow-capped mountains B. People depending on wild animals for food C. Poor, rocky, forest-covered soil near the sea D. Cool, damp climate; much low, wet ground E. Hot, dry climate, much soft, loose sand F. Hot, rainy climate; many raffia-palm trees G. Broad, level plains; rich soil; moderate summer rain ¹⁶

Another example of a matching exercise, involving cause-effect relationships, is the following set of items, taken from a standardized test in science. Actually, this represents a variation of the matching exercise, in which a series of multiple-choice items (134–138 below) are *classified* according to a key-list.

After each of statements 134–138, mark the letter designating the phrase below that will make the statement true.

- F if bacterial growth will be encouraged
 G if bacterial growth will not be affected
 H if bacterial growth will be decreased but will continue
 I if bacteria will be killed but spores will remain alive
 J if both bacteria and spores will be killed
- 134 Place bacteria in a refrigerator overnight
 135 Expose bacteria to a direct flame for 30 seconds
 136 Keep bacteria at a temperature of 98° Fahrenheit for 48 hours
 137 Put a solution of boric acid on bacteria
 138 Place bacteria on a medium of agar and beef broth¹⁷

The type of classification exercise given above is useful in many subject areas. For example, the key list might be a list of standard reference

¹⁶ Adapted from N. Theresa Wiedefeld and E. Curt Walther, *Wiedefeld-Walther Geography Test* (New York: Harcourt, Brace & World, Inc., 1931).

¹⁷ Reprinted with the permission of the California Test Bureau from Georgia Sachs Adams, William E. Keeley, and John A. Sexson, *California Tests in Social and Related Sciences*, Advanced, Form AA, Part III (Monterey, Calif.: California Test Bureau, 1954), items 134–138.

works; and the items, a series of study questions, to be classified according to the reference work most appropriate for answering each question.¹⁸

In a social studies unit in which the characteristics of democracy had been compared with those of other forms of government, a series of items could list several characteristics of, or practices under, democratic and totalitarian governments. Preceding this series of items, the following directions could be used:

After each item number on the answer sheet, blacken one lettered space to designate that the item is characteristic of the theory of

- A liberal democracy.
- B Communism.
- C Fascism.
- D both Communism and Fascism.
- E both liberal democracy and Communism.¹⁹

The following suggestions may help in the writing or revision of matching questions.

1. Do not include too large a number of items in either column. The number should probably vary from a minimum of 5 to a maximum of 12. The use of longer lists requires the student to spend too much time in hunting for the correct responses.
2. In any one question, do not mix items that are highly heterogeneous or dissimilar. For example, do not include in a single matching exercise items that require the matching of men and inventions with others that require matching of battles and dates.
3. The column of responses or options should include more alternatives than the column of questions or test items, in order to prevent the student from selecting the last response on the basis of elimination.
4. It is frequently advisable to allow certain items in the response column to be used more than once so as to reduce the effect of guessing. If this plan is used, the preceding suggestion becomes unnecessary.
5. If possible, the response column should contain shorter statements than the question column so that the student can scan the possible responses quickly.
6. The items in the response column should be arranged systematically if possible (names in alphabetical order, dates in chronological order, and the like). Note that in the example on page 343, responses F through J are in order of efficacy of bacterial control.
7. Double check to make sure that there is only one item in the response column that is the correct answer for each test item (unless the directions indicate that responses may be used more than once).
8. Avoid requiring the student to match parts of incomplete sentences because of the probability of introducing grammatical clues to the correct responses.
9. Be sure that a matching exercise appears on a single page of the test.

¹⁸ For an illustrative question of this type, see Georgia Sachs Adams and T. L. Torgerson, *Measurement and Evaluation for the Secondary School Teacher* (New York: Holt, Rinehart and Winston, Inc., 1956), p. 394.

¹⁹ Max D. Engelhart, "Exercise Writing in the Social Sciences," *Proceedings of the 1957 Invitational Conference on Testing Problems* (Princeton, N. J.: Educational Testing Service, 1958), p. 61.

A number of modifications of matching questions can be used. For example, students may be given a map or chart on which certain locations are assigned numbers or letters. These numbers or letters can then be matched with a list of cities, rivers, and the like.

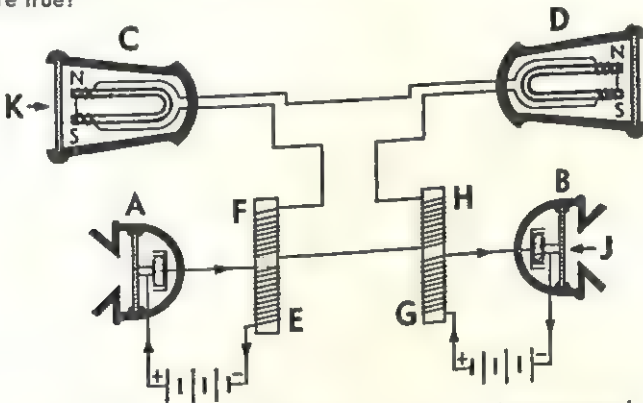
Situation- or Problem-solving Items

Situation- or problem-solving items deserve special mention because of their value in measuring student understandings, as contrasted with memorization or rote learning. A large part of the skill required in constructing tests of this type is involved in devising situations or problems that require the student to make interpretations from new data or to apply principles learned to new situations. The following examples are taken from the *California Tests in Social and Related Sciences*.

James wanted to be elected president of the school council. During the month before election he did the following things. Which one do you think was undemocratic?

- presented his plans concerning what he would do if elected president
- appealed to the city boys to "stick together" and not allow a country boy to win the office
- gave a talk in assembly asking the children to vote for him
- told the school that he would try to have more school picnics if he were elected²⁰

Examine the diagram of a two-way telephone set below. Which two of the following statements are true?



- In order to function properly, circuits AE and BG must be connected more directly.
- Diaphragm J vibrates when someone is talking into transmitter B.
- In receiver C, the purpose of electromagnet NS is to vibrate diaphragm K.
- Current cannot flow in AE unless someone is talking into it.²¹

²⁰ Reprinted with the permission of The California Test Bureau from Georgia Sachs Adams and John A. Sexson, *California Tests in Social and Related Sciences, Elementary, Part I* (Monterey, Calif.: California Test Bureau, 1953), Form AA, item 83.

²¹ Reprinted with the permission of The California Test Bureau from Georgia Sachs Adams and others, *California Tests in Social and Related Sciences, Advanced, Part III* (Monterey, Calif.: California Test Bureau, 1954), Form AA, item 41.

Problem-solving tests may be of the essay or completion, as well as the multiple-choice, type. For example, a map, graph, or table can be presented and students can be asked to draw generalizations from it. Such generalizations can later be used by the teacher in building a multiple-choice test on the same material.

Although situation-type exercises are difficult to construct, they are valuable in that they stimulate the functional use of facts and generalizations. A number of illustrations of situation- or problem-solving items are given in the next chapter, in which we present illustrative exercises for each category of the taxonomy.

Since situation- or problem-solving items may be of many types, it is impossible to make a list of specific suggestions for their construction. It is important that the situation contain an element of novelty without being so novel as to be unrealistic or confusing. The situation should be a challenging one to the students, and the alternative choices should all be plausible. The student should not be able to make the correct responses without careful reading of the question and critical thinking. Ordinarily, a good situation-type problem is one that would provide a good basis for class discussion. Before he attempts to write the alternative responses necessary for an objective test of this type, the teacher may wish to use a situation-type test as a basis for class discussion or for written responses by students.

Situation-type items are frequently used to measure the student's ability to apply principles learned in new situations. Hence, Dunning's step-by-step outline for developing tests on application of principles is relevant:

Step 1. Decide on the principle or principles to be tested. Criteria to be considered:

- a. Should be known principles but the situation in which the principles are to be applied should be new.
- b. Should involve significantly important principles.
- c. Should be pertinent to a problem or situation common to all students.
- d. Should be within the range of comprehension of all students.
- e. Should use only valid and reliable sources from which to draw data.
- f. Should be interesting to the students.

Step 2. Determine the phrasing of the problem situations so as to require the student in drawing his conclusion to do one of the following:

- a. Make a prediction.
- b. Choose a course of action.
- c. Offer an explanation for an observed phenomenon.
- d. Criticize a prediction or explanation made by others.

Step 3. Set up the problem situation in which the principle or principles selected operate. Present the problem to a class with directions to draw a conclusion or conclusions and give several supporting reasons for their answer.

Step 4. Edit the students' responses, selecting those that are most representative of their thinking. These will include conclusions and supporting reasons that are both acceptable and unacceptable.

Step 5. To the conclusions and reasons obtained from the students, the teacher now adds any others that he feels are necessary to cover the salient points. The total number of items should be at least 50 percent more than is desired in the final form to allow for elimination of poor items. The following list is a guide to the type of statements that can be used:

- a. True statements of principles and facts.
- b. False statements of principles and facts.
- c. Acceptable and unacceptable analogies.
- d. Appeal to acceptable or unacceptable authority.
- e. Ridicule.
- f. Assumes the conclusion.
- g. Teleological explanations.

Step 6. Submit test to other judges for criticisms. Revise test in view of criticisms.

Step 7. Administer test. Follow with thorough class discussion.

Step 8. Conduct an item analysis.

Step 9. In the light of steps 7 and 8, revise the test.²²

EVALUATING OBJECTIVE TEACHER-MADE TESTS

One of the best ways for a teacher to improve his skill in test construction is to evaluate his own tests—that is, to apply the generalizations developed in this chapter to specific test items. As a guide for such evaluation, the following checklist has been prepared. In applying this list to a teacher-made test, the teacher should check any relevant criticisms and note the specific test items or other characteristics that justify each check mark.

CHECKLIST FOR DISCOVERING FAULTY TEST ITEMS AND OTHER ERRORS IN THE CONSTRUCTION OF OBJECTIVE TESTS

Selection of Content

- 1. Failure to set up a table of specifications, or test blueprint, indicating the proportionate emphasis to be given to various objectives and content areas.
- 2. Failure to follow a test blueprint in the approximate distribution of test items.
- 3. Failure to emphasize the important facts and generalizations; test items emphasize mere details.

²² Gordon M. Dunning, "Evaluation of Critical Thinking," *Science Education*, vol. 38 (April 1954), pp. 191-193.

- 4. Introduction of material that is appropriate only for essay and discussion questions.

Other Factors Affecting Validity

- 5. Poor wording of items such as use of (a) bookish terms, (b) long involved phrases or statements, (c) vocabulary too advanced or unnecessarily technical.
- 6. Use of "specific determiners"—that is, clues afforded by phrasing, which tend to determine the student's response in the absence of knowledge. Specific examples are given in checklist item 23.
- 7. Ambiguous statements.
- 8. Failure to give all pertinent information necessary for student to choose answer.

Technical Make-up

- 9. Lack of directions.
- 10. Directions not specific or clear; failure to include sample exercise for unfamiliar type of item.
- 11. Use of items that help student to answer other items.

Other Factors Affecting Reliability

- 12. Too few questions to ensure an adequate sampling of the material tested.
- 13. Difficulty range not adequate for purpose.
- 14. Scoring not entirely objective—for example, scoring key does not include all possible answers to recall questions.

Physical Make-up

- 15. Mimeographing (or other means of duplication) poorly done.
- 16. Position of true-false questions, or position of answer-responses to multiple-choice questions, not randomized.
- 17. Faulty arrangement of the test items (questions crowded together; questions not grouped according to type; answer blanks scattered so that scoring is tedious; multiple-choice or matching question divided, that is, typed or printed on two different pages of test).

ERRORS PECULIAR TO A SPECIFIC TYPE OF TEST ITEM

True-False

- 18. Statements that are partly true and partly false.
- 19. Number of true statements excessively large. (Students who do not know answers tend to guess "true" more frequently than "false.")
- 20. Statements "lifted" from the text.
- 21. Use of true statements that are overlong. (A study conducted by a committee in one of the test-construction classes at the University of Wisconsin found that statements of 25 words or more were true in 80 percent of the cases).
- 22. Use of double negatives.
- 23. Use of "specific determiners." Items containing the following are more often false than true: *totally, entirely, exactly, completely, solely, fully, exclusively, very, perfectly, absolutely, all, always, no, none, not, nothing, only, alone, never*. Items containing the following are more often true

than false: *was one of, may, usually, generally, most commonly, as a rule, as a whole, even if, almost all, mainly, almost entirely, some, sometimes, often, frequently, several, many, probably, approximately, largely, ever.*

- 24. Crucial element of statement placed in a phrase or a subordinate clause.

Multiple-choice

- 25. More than one correct response (unless question is intended to have two or more correct responses).
—26. Distractors (false alternatives) that are not plausible.
—27. Responses too long, involving needless repetition (which could be reduced by rewording stem of item).
—28. Correct responses longer or more cautiously worded than others.
—29. Responses that are grammatically inconsistent with stem.
—30. Clues that help the unprepared student to select correct answer or eliminate one or more distractors.

Matching

- 31. Inconsistent or heterogeneous material within a column.
—32. Too few (less than 5) or too many (more than 15) items per question.
—33. Column of alternative responses not alphabetized or arranged in some other suitable systematic manner.
—34. Same number of items in both columns.
—35. "Specific determiners": identical elements, illogical statements, grammatical clues.

Completion and Simple Recall

- 36. Lack objectivity—that is, too many possible answers.
—37. Blanks not uniformly arranged for easy scoring.
—38. Too many blanks per item.
—39. Item consisting of a sentence from a text with one or two words omitted.
—40. Clues that help the unprepared student (such as number of blanks or blanks of varying length).
—41. Items requiring recall of trivial information.

PREPARING A TEACHER-MADE TEST FOR USE

Up to this point, we have been concerned with the planning of a test blueprint and the writing of different types of items. These are crucially important topics. The quality of a test is determined largely by the quality of its items. Careful attention to the editing of items, the writing of directions for students, the arrangement of items within the test, and the design of answer sheets and scoring keys can further improve test quality.

Editing Items

Test items should be written well in advance of the time that they are needed. The teacher who hastily reviews items he has just composed may

fail to catch ambiguities, for he reads into the item what he intends to communicate. Ideally, another teacher in the subject area should review the items, criticizing them and also indicating what he considers to be the correct responses. In requesting such assistance from a colleague, the teacher can emphasize that he would much rather revise or eliminate items than face the ill feelings that inevitably develop among students when test questions are ambiguous. In addition to revising ambiguous questions at this time, the teacher should try to substitute improved options for any which are not plausible, revise the stem and/or options for any items incorrectly keyed by a colleague, and eliminate (or revise) any items that appear to be extremely easy or difficult for the group.

The teacher should have written at least 20–30 percent more items than are needed so that (1) he can afford to discard items that cannot be satisfactorily revised and (2) he has some “elbow room” in adjusting the number of items to his table of specifications.

Grouping Test Items

After the items have been reviewed and the necessary changes made, they should be arranged for typing. If items are on cards, their arrangement in any desired sequence can easily be accomplished. There are three possible bases for grouping: (1) type of item (that is, true-false, recall, and the like), (2) difficulty of item, and (3) content. The chief reasons for the first type of grouping are (1) so that instructions for answering can be carried throughout a set of items, and (2) so that standard answer sheets can be used. This first basis for grouping is almost uniformly used.

Within each type of item (true-false, recall, and the like), the teacher can choose to group either by difficulty or content. In a very long test, he might do both. Arranging items in approximate order of difficulty is especially important if there is insufficient time for all students to complete the test; thus the slow-working student does not waste time on questions that are “beyond him,” nor does he get discouraged early in the test.

If a class is fairly homogeneous with respect to student ability, and adequate testing time is allowed, arranging items by approximate difficulty may be less important than arranging items by content. Arranging items according to content makes the test have higher face validity for the examinee and may aid the teacher in judging the areas in which review or class discussion is desirable.

Writing Directions for Examinees

For high school or college students who are test-wise, directions for the usual true-false or multiple-choice items may seem superfluous. However, it is far better to include instructions than to have some students

make mistakes in their method of indicating replies, or to have class time wasted by questions on test procedure.

The following sample directions may be used or adapted if a home-made answer sheet (similar to Figure 10.2) is used. If students indicate their answers on the test itself, obvious changes can be made to make the directions simpler and briefer.

True-False Items

Read each of the following statements. Indicate your answers on the separate answer sheet. *Make no marks on the test.* If a statement is true, cross out the A on your answer sheet. If a statement (or any part of a statement) is false, cross out the B.

Multiple-Choice Items

Read each of the following items. Indicate your answers on the answer sheet. *Make no marks on the test.* Decide which of the options or choices *best* completes the statement or answers the question. If you think that choice A is the best answer to Item 1, cross out the A in the row after No. 1 on your answer sheet; if you think choice B is best, cross out the B, and the like.

Name _____		Period _____	
(last name)		(first name)	
Course _____		Date _____	
Title of examination _____			
<p><i>Directions:</i> Fill in all the information requested above. <i>Make no marks on your test.</i> Read the directions on the test carefully, and follow directions exactly. For each multiple-choice item, mark your choice for the correct answer by crossing out the letter that corresponds to your choice. For True-False items, indicate your choice of True by crossing out the A; or your choice of False by crossing out the B.</p>			
Item No.	Item No.	Item No.	Item No.
(T) (F)	(T) (F)	(T) (F)	(T) (F)
1. A B C D E	21. A B C D E	41. A B C D E	61. A B C D E
2. A B C D E	22. A B C D E	42. A B C D E	62. A B C D E
3. A B C D E	23. A B C D E	43. A B C D E	63. A B C D E
4. A B C D E	24. A B C D E	44. A B C D E	64. A B C D E
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
20.	40.	60.	80.

Fig. 10.2 An Illustrative Mimeographed Answer Sheet (to be used with a lay-over scoring stencil).

Sample directions for modified true-false items have been given on page 338 and for a matching item on page 342. In Chapter 11 a number of items with which students might not be familiar are included. The directions given for these items will be helpful to teachers in devising instructions for similar types of items.

If a test is being given with a liberal time allowance, as is true for most teacher-made tests, students can be asked to answer every question. Under such circumstances, simple number-right scores rank students in the same order as scores corrected for guessing. Hence, unless the teacher wishes to discourage guessing, no correction formula need be used.

If the teacher wants to use a "correction-for-guessing" formula, tests should be scored for both rights and wrongs. The usual formula for a corrected score is

$$\text{Number right} - \frac{\text{Number wrong}}{n - 1}$$

where n is the number of options. For true-false questions, substitution of 2 for n gives a corrected score of $R-W$. For five-choice questions, one-fourth of the number wrong is subtracted from the number right. Since this formula tends to penalize the overcautious student,²³ the test instructions should encourage students to utilize their partial knowledge. The following advice to students taking College Entrance Board Examinations is much better than a warning not to guess.

Many candidates are unsure of the wisdom of guessing at the answers to questions about which they are uncertain. For each of the College Board Achievement Tests, . . . a percentage of the wrong answers is subtracted from the number of right answers as a correction for haphazard guessing. It is highly unlikely, therefore, that blind guessing will improve your score significantly; it may very well lower it, and it does take time. Often, however, you

²³ The formula for "correction for guessing" is based on the assumption that the student's chance of picking the correct response on items he doesn't know is $\frac{1}{n}$. That is, if there are n choices, the chance probability of a correct guess is $\frac{1}{n}$ and of an unlucky or wrong guess, $\frac{n-1}{n}$. For every $n-1$ wrong guesses, we expect one correct guess. Hence, to correct for guessing, one subtracts $\frac{\text{Number wrong}}{n-1}$ from the score (number right). This formula is based on the faulty assumption that items can be categorized neatly into those the student knows and those he does not know. When a correction formula is used, the student who is willing to gamble has an advantage over the student who omits all items about which he is doubtful. The student who is willing to gamble capitalizes on his partial knowledge as well as weaknesses in test items; that is, if a student can easily eliminate two of the five options as illogical, his chance of getting the item right is higher than the formula assumes.

will not be sure of a correct answer but will have some knowledge of the question. If you can eliminate one or more of the answer choices as definitely wrong, it will be to your advantage to answer the question even though you must make a guess as to which of the remaining answers is correct.²⁴

Instructions should be given regarding the policy students should follow when they are uncertain about answers. These instructions should be included as part of the regular set of directions, rather than given orally, or in response to individual students who come up to inquire about it.

Physical Make-up of the Test

A common mistake in teacher-made tests is to try to crowd too much material on a page. As a rule, each option for a multiple-choice item should be typed on a separate line. If the responses are very brief, two columns can be used, as on page 371. The items that refer to a map, table, or chart should be on the same page as that material. A matching exercise or multiple-choice item should not be divided between two pages.

If a separate answer sheet is not being used, the test itself should be designed to facilitate scoring. Most students in the fourth grade and above can copy their answers in an answer column so that the teacher can lay his scoring key beside this column. If the teacher wishes to use a cut-out, lay-over scoring stencil, students can encircle the words "true" or "false" (or the letters associated with options) that are typed in an answer column. Question numbers should be repeated in the answer column.

If students are being asked to encircle the letter of an option, it is best not to repeat the same letters for all questions but to use the following plan, which helps younger students to keep their place.

1. A B C D E
2. F G H I J
3. A B C D E

Needless to say, the options have to be lettered correspondingly in the test items.

If separate answer sheets are used for teacher-made tests given in elementary schools, the test should be typed first, and then the answer sheet should be typed so that the answer spaces are right in line with the end of each test question, rather than uniformly spaced as in Figure 10.2. With the use of this procedure, plus the alternation of letter patterns suggested above, separate answer sheets can be used with many fourth-, fifth-,

²⁴ *A Description of the College Board Achievement Tests* (Princeton, N. J.: Educational Testing Service, 1962), p. 17.

and sixth-grade classes. Answer sheets have advantages in ease of scoring, economy of reusing tests, and use in item analysis, discussed in the next chapter section.

STATISTICAL ANALYSIS OF TEST RESULTS

The authors of standardized tests always write many more test items than are needed and decide, on the basis of a preliminary tryout, which items should be revised or eliminated. They study the *difficulty* of each item so that they can select items which meet desired standards with respect to difficulty. They also study the relationship of each item with the *criterion*. Reference to the summary tables of Chapter 4 will reveal that if a predictor test is being devised, the author's goal is to select items that show a fairly high relationship with an external criterion. If an achievement test is being developed, the criterion is usually total test score. That is, on an achievement test, efficiency of measurement can be improved by discarding or revising items that do not correlate well with total score, or do not *discriminate* between high-achieving and low-achieving students (on the test as a whole). Such a study of the difficulty and discrimination value of items is called an item analysis.

Before we consider item analysis procedures that are more suitable for teachers, we will discuss a procedure frequently used by authors of standardized tests, that is, finding for each item its *bi-serial r* with total test score. *Bi-serial r* is a special type of correlation coefficient, computed when *one* of the variables is dichotomous (that is, has only two possible values). The score for an item is either 1 or 0, but the total test score can have many values. Hence this type of coefficient is appropriate.

Since *bi-serial r* is used so frequently in test construction, short-cut procedures have been devised. One can merely compute the proportion of students succeeding on an item in the high-scoring group and the corresponding value for the low-scoring group;²⁵ then one looks these values up in a table and finds the value of *bi-serial r* .²⁶ The test author then selects items with fairly high *bi-serial r* 's that also provide a satisfactory distribution of item difficulties and (in achievement tests) a satisfactory

²⁵ For machine-scored tests, the number of students succeeding on each item is easily obtained by the use of an item-count attachment to the test-scoring machine. Item data for one hundred test papers can be counted by machine in ten minutes. An illustration of a graphic item count record is given in Figure 14.4.

²⁶ Chung-Teh-Fan, *Item Analysis Table* (Princeton, N.J.: Educational Testing Service, 1952). A procedure for approximating the *bi-serial r* when item analysis data for the high-achieving and low-achieving halves of the class are compared is given in Figure 10.3 of this textbook.

distribution with respect to content areas and objectives. An average *bi-serial r* of .40 is considered adequate, while .50 is exceptionally high.

In selecting the high-scoring and low-scoring groups for comparison, the customary procedure is to select the highest and lowest 27 percent of the group, because this percentage has been found to work most effectively when results are analyzed for a fairly large group of students. However, teachers usually work with groups of 50 students or fewer; hence it is best to compare results for the high-achieving and low-achieving *halves* of the group, because the results have greater reliability (that is, tend to vary less from one sampling of students to another).

After considerable experimentation, Diederich²⁷ has developed procedures for having students help in item analysis by a show of hands. He reports that these procedures can be used over a wide grade span. Even fourth-grade classes do not find it too difficult, while college students do not resent the use of class time. In fact this use of class time can be justified in that the remainder of the period can be used in discussion of the questions most frequently missed. The procedures may be summarized as follows:

1. After the papers have been scored, the teacher records the distribution of scores on the chalkboard. If tests have just been scored in class, the teacher can call off successive scores, beginning with the highest, while a student assistant counts the number of hands raised for each score.
2. The teacher then quickly finds the median or midscore. All papers are collected and quickly sorted as above, at, or below the midscore. Papers in the high-scoring half of the class are passed to one side of the room (say the left side), the low-scoring papers to the other side. One-half of the papers in the median interval are assigned to each side. Thus, each student has a paper, and presumably no student has his own. (If the total number of students in the class is an odd number, one of the papers in the median interval is excluded.)
3. For each item, the teacher records on the board (and students record following each item on their papers) four figures, which are labeled and defined as follows:

H	L	H + L	H - L
(The number of highs who got the item right)	(The number of lows who got the item right)	<i>Success</i> (The total number who got the item right)	<i>Discrimination</i> High-low difference (how many more highs than lows got the item right)

For example, the teacher calls "item 1." Everyone whose paper has the item right raises his hand. A student assigned to count the highs calls out the

²⁷ Paul B. Diederich, "Item Analysis," in *Short-Cut Statistics for Teacher-Made Tests*, Evaluation and Advisory Service Series No. 5 (Princeton, N.J.: Educational Testing Service, 1960), pp. 1-10.

number of upraised hands in the high section; then the counter for the lows calls out the number of hands in his group. The teacher (or student score-keeper) calls out the total and then the difference. For example, if there were fifteen highs and ten lows for item 1, the sequence would be: "item 1, hands [*pause for counting*] 15-10-25-5; item 2. . ."

The items with the highest success values are the easiest for this group. If the teacher wishes, he can easily change these to percentages for his own use by multiplying by the constant, $\frac{100}{n}$. For example, if there were

40 in the class, each success value would be multiplied by $\frac{100}{40}$ or 2.5. The success value for this illustrative item would be $2.5 \times (25)$ or 63 percent. If the test is designed to rank pupils (as a basis for grading), inclusion of too many easy items is not good use of testing and scoring time. On the other hand, very difficult items should be examined for ambiguity, especially if failed by a large percentage of the high group.

The minimum high-low difference corresponding to the standards of professional test construction would be 10 percent of the group, or four pupils in a class of 40. However, because of the variation from class to class in student performance on items, we would want to examine, rather than discard, items with high-low differences smaller than this. An item can frequently be improved by substituting options that will attract more choices from poorly prepared students. The teacher must not routinely discard items that fail to achieve a specified discrimination value without considering the effect on the test's balance of items by content areas (according to his own table of specifications).

If time permits, the show-of-hands procedure can also be used for the second stage of item analysis, that is, the number of students choosing each option. If there is not time, this stage (which involves only those items that showed low or negative discrimination values)²⁸ can be completed by the teacher at home.

Which of the following contains the most heat?

	Highs	Lows
a. A red-hot branding iron	0	10
b. A lake full of water	20	8
c. A car engine that has been driven in the heat	0	0
d. A pail of boiling water ²⁹	0	2

²⁸ Negative discrimination values are obtained when more low-scoring than high-scoring students get an item correct.

²⁹ Adapted from Gilbert Sax, *The Construction and Analysis of Educational and Psychological Tests* (Madison, Wis.: College Printing and Typing Company, Inc., 1962), p. 47.

Obviously option *a* is a good distractor; however, option *c* has drawn no choices. Since the terms "red-hot" in option *a* and "boiling" in option *d* may have drawn choices to those options, one might substitute for option *c* such a distractor as "A barrel of hot tar." This type of item analysis may also help the teacher in group diagnosis, that is, in his understanding of misconceptions held by members of the class. Sometimes distractors can be written to represent different types of common errors and misunderstandings; for example, the distractors in arithmetic problems can represent types of procedural errors frequently made by students.

If the use of class time in item analysis cannot be justified, or if the test is a final examination, which is scored after the last class session, item analysis by machine can be used; or the teacher can use a short-cut procedure devised by Katz.³⁰

TEACHER COOPERATION IN TEST DEVELOPMENT

Progress toward improving teacher-made tests is greatly facilitated if the teachers of multiple-section courses in a school or school district cooperate in the development of tests, or at least a file of test items, in each subject field. There are many advantages in teacher cooperation in the development of tests. Far better tests are likely to result from cooperative discussion of the objectives of a course and the types of items that might be effectively used in measuring student growth toward such objectives. Then, if the work can be divided on the basis of subject areas or objectives, each teacher can concentrate on writing and revising a smaller number of test items than he would need for an entire end-of-course examination. When the test items have been written and copies of all items distributed to the teachers, each teacher should "take the test." Differences of opinion regarding the keying of items will help to discover ambiguous questions before they are actually used.

Another significant advantage of teacher cooperation is that the results of item analysis can be pooled. The item difficulty and discrimination values obtained through pooling results from several classes are much more reliable than could be obtained by one teacher. If teachers do not wish to have a common departmental test, they can select their own items from a master file and record their item analysis data for each item, as shown in Figure 10.3.

³⁰ Martin Katz, "Improving Classroom Tests by Means of Item Analysis," *The Clearing House*, vol. 35 (January 1961), pp. 265-269.

U S HISTORY
TERRITORIAL
EXPANSION

1.12

Item: Texas became part of the United States by

- a. purchase from France
- b. purchase from Spain
- c. treaty with Spain after the Spanish-American War
- d. request of the people of Texas

Difficulty (% success)	70%,	60%,	75%	
Estimated bi-serial r^a	.39,	.42,	.36	
Number choosing each option:		a	b	c
High-scoring half			2	18
Low-scoring half	3	3	1	13

Fig. 10.3 An Item Card for a Departmental Item File

^a If the performance of students in the high-scoring and low-scoring halves of the class (above and below the median on total test score) is compared, the bi-serial r for items of moderate difficulty is approximately equal to three times the high-low difference (expressed as proportion of group). For example, if the high-low difference is 10 percent or .10, the bi-serial r , or discrimination index, would be $3 \times .10$ or .30. If the high-low difference is 15 percent or .15, the index would be $3 \times .15$ or .45. This estimate is approximately correct for items which 20-80 percent of the students answer correctly. For very easy or very difficult items, this approximation underestimates the index. For an item on which more than 80 percent succeed, a high-low difference of 5 percent is acceptable; whereas for other items, the difference should be at least 10 percent, with the corresponding index being at least .30. Because of the sampling error involved in studying small groups, items with low discrimination indexes on one trial may have an index above .30 with another class. Ordinarily items with low indexes should be revised, for example, by substituting a better distractor for one which is not attracting the choices of low-scoring students. The methods of making an item analysis of a test by a "show of hands" in class, together with short-cut procedures for analyzing the data, are given in Paul B. Diederich, *Short-Cut Statistics for Teacher-Made Tests*, Evaluation and Advisory Service Series No. 5 (Princeton, N. J.: Educational Testing Service, 1960). Single copies are available free on request to the Educational Testing Service, 20 Nassau Street, Princeton, New Jersey.

An intermediate approach (intermediate between free-lancing and a common departmental test) could be used. Such an approach (that is, the use of an "anchor test") would allow for variations in course content, which may not only be legitimate but desirable; yet it would achieve some of the advantages of a common test. Teachers can usually agree on at

least one-half of the items of their end-of-course examination, which could then constitute an anchor test. The other half of the items could be selected or devised by the individual teacher. A tabulation of scores for each class on the anchor test would help teachers to modify their grading distributions from class to class (in terms of any significant deviation between a class distribution of anchor test scores and that for all students combined).

For a file of test items to be of maximum usefulness, each card should contain not only the item itself and the keyed answer but (1) classification of the item by content area and by objective and (2) the *cumulated* results of experience with the item. An illustrative test item card is shown in Figure 10.3. The notation in the upper left-hand corner indicates the subject, and content area within the subject; while the number in the upper right-hand corner indicates that the item involves "1.12 Knowledge of Specific Facts," a subcategory of the taxonomy discussed in the next chapter.

PROVIDING LEADERSHIP IN THE DEVELOPMENT OF TEACHER-MADE TESTS AND OTHER AIDS TO EVALUATION

Since it is obvious that standardized tests can do only a portion of the job of assessing student growth, administrators and supervisors have the further responsibility of helping teachers learn to develop better teacher-made tests.

Ebel has listed seven serious errors that teachers frequently make in their evaluation of pupils' educational attainments.

First, teachers tend to rely too much on their own subjective, but presumably absolute, standards . . . the unreliability of subjective judgments have been demonstrated over and over again. Yet few teachers have been persuaded to use pooled judgments in cooperative test construction . . . or to recognize their inevitable use of relative standards in evaluating student attainments. . . .

Second, teachers tend to put off test preparation to the last minute. . . .

Third, many teachers use tests which are too poorly planned, too short, or too inefficient . . . to sample adequately all the essential knowledge and abilities in the area of educational attainment covered by the tests. . . .

Fourth, teachers often place too much emphasis on trivial . . . details . . . to the neglect of basic principles, understandings, and applications. . . .

Fifth, teachers often write questions, both essay and objective, whose effectiveness is lowered by ambiguity, or by irrelevant clues to the correct response. Too seldom do they seek even one independent review of their questions by a competent colleague.

Sixth, many teachers overlook, or underestimate, the magnitude of the sampling errors which affect test scores. . . . Differences as small as one score unit are considered to reflect significant differences in attainment.

Seventh, most teachers fail to test the effectiveness of their tests by even a simple statistical analysis of the results from the test. . . . There is no better way to develop skill in testing than to analyze systematically the results of previous efforts.³¹

Preparation of teachers for their inescapable responsibility of evaluating student achievement should be begun at the preservice level. However, proficiency in the practical art of educational measurement requires considerable "learning by doing" on the job.

Ebel has observed hundreds of teachers at work on committees charged with developing new tests for the Educational Testing Service; he has studied the reactions of these teachers and the test specialists who work with them. On the basis of his experience in seeing teachers grow in their interest and competence in test development, Ebel³² contends that highly effective in-service education in measurement results when a group of teachers work together on (1) developing specifications for a test they all need, (2) writing items and revising each other's items, (3) preparing the test for use, (4) trying it out, and (5) analyzing the results.

At several crucial points in their cooperative work (probably at intervals of 4-6 weeks), a specialist in test construction could be of considerable assistance to such a group of teachers. If the school district did not have such a specialist on its staff, periodic visits by a consultant would be a justifiable expenditure. Teacher concern about the development of more adequate tests inevitably leads to concern about the appropriateness and clarity of objectives, the quality and relevance of educational experiences, and the effectiveness of the teaching process. Such concern is an essential prerequisite to improvement in curriculum and instruction.

The Cooperative Test Division of Educational Testing Service has prepared a series of color filmstrips, with synchronized sound, for use in a teacher workshop on "Making Your Own Tests." Each filmstrip (running 25 minutes each) is focused on a major step in the process, for example, "planning," "construction," and "analysis" of classroom tests. Included in the kit are 28 ditto masters that summarize and amplify the major points of each film.³³ Each school district can duplicate the number of copies it needs for its own use. The Cooperative Test Division has also produced five 15-minute, 16mm. sound films for use in training programs for counselors and administrators (or with teachers if a qualified discussion leader is present). The films are *not* suitable for use with lay

³¹ Robert L. Ebel, "Improving the Competence of Teachers in Educational Measurement," *The Clearing House*, vol. 36 (October 1961), pp. 67-71.

³² *Ibid.*, pp. 69-70.

³³ Further information concerning this kit may be obtained from the Director of Educational Relations, Cooperative Test Division, Educational Testing Service, Princeton, N.J.

groups. The titles are "Selecting an Achievement Test," "Administering a Testing Program," "Interpreting Test Results Realistically," "Using Test Results," and "The Public Relations of Testing." These films may be rented or purchased.

SUMMARY STATEMENT

If a teacher-made test is to be used to grade students, or rank them with respect to total score, the test should (1) be based on a representative sampling of the content studied, (2) be based on a representative sampling of the abilities or skills emphasized in the course, (3) contain a sufficient number of questions so as to have adequate reliability, and (4) include items covering a wide range of difficulty. If a test is designed to serve special instructional purposes rather than to rank students on the basis of scores, the criteria listed above are less applicable. For such tests, the following criteria are more significant: (1) Does the test elicit from the students the desired type of mental processes? (2) Does the test encourage the development of desirable study habits? (3) Does the test lead to improved instructional practice? (4) Does the test foster wholesome relationships between teachers and students?

Although essay examinations have the advantages of being easily prepared, reducing the amount of guessing, and stimulating superior study methods, they favor students with linguistic ability and have relatively low reliability and validity. If skillfully constructed, however, they may give the student opportunity to demonstrate his ability to select and organize his learnings and may provide the teacher with a basis for diagnosing errors in interpretation and concept formation.

The use of objective or selection-type test questions makes it possible to test a large sampling of learnings in a relatively short testing time. This extensiveness of sampling and the objectivity of scoring contribute to the relatively higher reliability of the objective test. Objectivity of scoring also reduces scoring time and introduces the possibility of test scoring by students or clerical workers, or by a test-scoring machine. Specific suggestions for the construction of each type of item were given, as well as a summary checklist for discovering faulty test items and other errors in test construction.

Practical suggestions were given for editing test items, writing test directions, developing answer sheets, and making an item analysis of student responses to test items.

SELECTED REFERENCES

- BEARD, RICHARD L., "Techniques the Teacher May Use in Constructing Tests," *High School Journal*, vol. 36 (January 1953), pp. 101-106.
- BLOOM, BENJAMIN S., ed., *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain*. New York: David McKay Company, Inc., 1956.
- CONRAD, HERBERT S., "The Experimental Tryout of Test Materials," in E. F. Lindquist, ed., *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, Chapter 8.
- DAVIS, FREDERICK B., "Item Analysis in Relation to Educational and Psychological Testing," *Psychological Bulletin*, vol. 49 (March 1952), pp. 97-121.

- DIEDERICH, PAUL B., "Making and Using Tests," *English Journal*, vol. 44, (March 1955), pp. 135-140, 151.
- EBEL, ROBERT L., "Procedures for the Analysis of Classroom Tests," *Educational and Psychological Measurement*, vol. 14 (Summer 1954), pp. 352-364.
- ENGELHART, MAX D., "How Teachers Can Improve Their Tests," *Educational and Psychological Measurement*, vol. 4 (Summer 1944), pp. 109-124.
- FRENCH, WILL, AND ASSOCIATES, *Behavioral Goals of General Education in High School*. New York: Russell Sage Foundation, 1957.
- FURST, EDWARD J., *Constructing Evaluation Instruments*. New York: David McKay Company, Inc., 1958.
- GERBERICH, J. RAYMOND, *Specimen Objective Test Items*. New York: David McKay Company, Inc., 1956, Parts I, II, and III.
- LENNON, ROGER T., "Testing: Bond or Barrier between Pupil and Teacher," *Education*, vol. 75 (September 1954), pp. 38-42.
- LINDQUIST, E. F., "Preliminary Considerations in Objective Test Construction," *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, Chapter 5.
- NOLL, VICTOR H., "Objectives as the Basis of All Good Measurement," *Introduction to Educational Measurement*. Boston: Houghton Mifflin Company, 1957, pp. 90-107.
- VAUGHN, K. W., "Planning the Objective Test," in E. F. Lindquist, ed., *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, Chapter 6.
- WEITZMAN, ELLIS, AND WALTER J. MCNAMARA, *Constructing Classroom Examinations*. Chicago: Science Research Associates, Inc., 1949, Chapters 2-4.
- WOOD, DOROTHY ADKINS, *Test Construction: Development and Interpretation of Achievement Tests*. Columbus, Ohio: Charles E. Merrill Books, Inc., 1960.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. Summarize the advantages and limitations of essay-type tests in your subject field or grade level.
2. Summarize the advantages and limitations of objective tests in your subject field or grade level.
3. Construct an essay-type examination consisting of five questions, and prepare an objective scoring key.
4. Prepare five true-false and five multiple-choice items on the contents of this chapter, following the suggestions given for the construction of these types of items.
5. Construct five matching-type items on the contents of this chapter, following the suggestions given. Have the items evaluated by two or more of your classmates.
6. Obtain an informal, objective teacher-made examination and evaluate it, using the checklist of errors given in this chapter.
7. Construct an informal objective test of 50 items, and have it evaluated by a classmate. Revise the items found to be faulty.
8. What are the advantages of maintaining a file of test items in your subject field? Of working cooperatively with other teachers to maintain a departmental file of this type? What facts should be recorded for each item?

The Taxonomy of Educational Objectives and Test Items Illustrative of Its Major Categories

A great step forward in recognizing the complexity of educational objectives and the difficulties involved in measuring student achievement was taken when Bloom and several professional associates worked together over a period of years in developing a taxonomy of educational objectives, under which educational goals and test items in the cognitive areas could be classified.¹

The major categories of this taxonomy, and each of its various subcategories, are summarized in Table 11.1, together with illustrative objectives, classified under each subcategory. A study of this table reveals that these objectives are classified under six main headings, which represent increasing degrees of complexity. The objectives in each major class make use of, and are dependent on, the goals of the classes lower in the hierarchy; for example, the cognitive abilities classifiable under "3.00 Application" are dependent on the abilities classifiable under "2.00 Comprehension," which in turn are dependent on those under "1.00 Knowledge." As an aid to students in understanding the taxonomy, we have included in this table one or more objectives of a course in measurement for each of the 20 subcategories.

The six sections of this chapter are devoted to an explanation of each of the six major categories of the taxonomy and to the presentation of test items illustrative of each category. Not only will these test items serve to illustrate the taxonomy but they will provide examples of item writing for several subject fields. For additional suggestions for item writing in various subject fields, the student is referred to a list of selected references, classified by subject field, listed in the bibliography for this chapter.

¹ Benjamin S. Bloom, *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain* (New York: David McKay Company, 1956). A similar volume devoted to the development of a taxonomy for objectives in the affective domain is nearing completion.

Table 11.1

Examples of Course-of-Study Objectives Classified under the Headings
of the Taxonomy of Educational Objectives: The Cognitive Domain

-
- 1.00 KNOWLEDGE^a (Remembering facts, terms, and principles in the form that they were learned)
- 1.10 *Knowledge of Specifics*
- 1.11 *Knowledge of terminology*
 Know the terms used in the study of elementary chemistry
 Know the terms used for different types of converted scores
- 1.12 *Knowledge of specific facts*
 Know the physical and chemical properties of common elements
 Know the major sources of information on published tests
- 1.20 *Knowledge of Ways and Means of Dealing with Specifics*
- 1.21 *Knowledge of conventions*
 Know the ways in which symbols are used in writing equations in chemistry
 Know the conventional procedures followed in tallying data in a frequency distribution
- 1.22 *Knowledge of trends and sequences*
 Develop a basic knowledge of the evolutionary development of man
 Know the major trends in the development of achievement testing
- 1.23 *Knowledge of classifications and categories*
 Know the major branches of biological science
 Know the major classifications of the "Taxonomy of Educational Objectives: Cognitive Domain"
 Know the four different types of number systems
- 1.24 *Knowledge of criteria*
 Know the criteria by which the nutritive value of a meal can be judged
 Know the major criteria to be used in evaluating tests for a specific purpose
- 1.25 *Knowledge of methodology*
 Know the methods for estimating the size of distant stars
 Know the methods used in obtaining norming samples that are representative of high school students
- 1.30 *Knowledge of the Universals and Abstractions in a Field*
- 1.31 *Knowledge of principles and generalizations*
 Know the biological laws of reproduction and heredity
 Know the principles involved in comparing reliability coefficients computed on different groups

Table 11.1 (Continued)

Examples of Course-of-Study Objectives Classified under the Headings
of the Taxonomy of Educational Objectives: The Cognitive Domain

1.32 Knowledge of theories and structures

Know a relatively complete formulation of the theory of evolution

Know a relatively complete formulation of the four types of validity
and their relationships to the types of judgment to be made

2.00 COMPREHENSION^b (Understanding material studied without necessarily relating it to other material)

2.10 Translation (from one set of symbols to another)

Can prepare graphical representations of physical phenomena, or of observed and recorded data

Can translate normalized standard scores into percentiles

Can read percentiles, stanines, or normalized standard scores from an Otis Normal Percentile Chart

Can translate a formula into a verbal explanation or vice versa

2.20 Interpretation (summarization or explanation)

Can distinguish among warranted, unwarranted, or contradicted conclusions drawn from a body of data

Can read the section on Reliability in a test manual and summarize the data in his own words

2.30 Extrapolation (extension of trends beyond data given)

Can interpolate when there are gaps in the data

Can estimate the probable reliability coefficient for his own group on the basis of reliability data published in the test manual

3.00 APPLICATION^c (Using generalizations or other abstractions appropriately in concrete situations)

Can predict the probable effect of a change in a factor on a biological situation previously at equilibrium

Can select, on the basis of a table of intercorrelations, the best pair of tests for the prediction of success in a specific school or job situation

4.00 ANALYSIS^d

4.10 Analysis of Elements

Can distinguish a conclusion from the statements that support it

Can distinguish between facts and interpretations in reading anecdotal records or other case study materials

4.20 Analysis of Relationships

Can check the consistency of hypotheses with given information and assumptions

Can recognize which methods of studying reliability are relevant to a particular use of a test

Table 11.1 (Continued)

Examples of Course-of-Study Objectives Classified under the Headings of the Taxonomy of Educational Objectives: The Cognitive Domain

4.30 *Analysis of Organizational Principles*

- Can recognize the general structure of a musical composition
- Can recognize the techniques used in persuasive materials, such as advertising and propaganda
- Can recognize the differences among the four types of validity with respect to their emphasis on the criterion

5.00 SYNTHESIS^a

5.10 *Production of a Unique Communication*

- Can write simple musical compositions
- Can tell a personal experience effectively
- Can devise test items classifiable under each major division of the taxonomy

5.20 *Production of a Plan or Proposed Set of Operations*

- Can propose experiments for testing hypotheses
- Can design simple machine tools to perform specified operations
- Can set up a table of specifications for a test on a specific unit of work
- Can devise a plan for the local validation of a test for a specific purpose

5.30 *Derivation of a Set of Abstract Relations*

- Can discover and formulate generalizations in mathematics
- Can discover and formulate an appropriate set of categories to use in summarizing student responses to a free-response question, such as "What do you like best about your school"

6.00 EVALUATION¹ (Judging the value of material for a specified purpose)

6.10 *Judgments in Terms of Internal Evidence* (for example, logical consistency)

- Can indicate logical fallacies in arguments
- Can evaluate the adequacy with which research data in a test manual support statements made about the value of the test for certain purposes

6.11 *Judgments in Terms of External Criteria*

- Can make a comparative appraisal of two or more tests for a specific purpose on the basis of the criteria presented in Part I of the text and consultation of expert opinion in the *Buros Mental Measurements Yearbooks*
-

Source: The headings of this outline and some of the illustrative objectives are taken verbatim from Benjamin S. Bloom, ed., *Taxonomy of Educational Objectives: Handbook I, Cognitive Domain* (New York: David McKay Company, Inc., 1956), Part II. Note that the taxonomy is arranged in a hierarchy, that is, each classification within it utilizes the skills and abilities that are lower in the classification order; for example, "Application" requires both "Comprehension" and ability to recall "Knowledge."

^a In the taxonomy, the term "knowledge" is defined as including "those behaviors and test situations that emphasize the remembering, either by recognition or recall, of ideas, material,

Table 11.1 (Continued)

Examples of Course-of-Study Objectives Classified under the Headings
of the Taxonomy of Educational Objectives: The Cognitive Domain

or phenomena. The behavior expected of the student in the recall situation is very similar to the behavior he was expected to have during the original learning situation." [Italics added.] *Ibid.*, p. 62.

^b In the taxonomy, the term "comprehension" is used to represent "an understanding of the literal message contained in a communication without necessarily relating it to other material. Three types of comprehension behavior are considered. . . . The first is *translation*, which means that an individual can put a communication into another language, into other terms, or into another form of communication. . . . The second type of behavior is *interpretation*, which involves dealing with a communication as a configuration of ideas whose comprehension may require the reordering of ideas into a new configuration in the mind of the individual. . . . The third type . . . is *extrapolation*. It includes the making of estimates or predictions based on understanding of the trends, tendencies, or conditions described in the communication." *Ibid.*, pp. 89-90.

^c While comprehension requires that a student understand a generalization or other abstraction well enough to illustrate its use, *application* requires that the student select the appropriate generalization or concept and use it in a situation in which no mode of solution is suggested. To test application, a test situation must either be novel or must contain novel elements as compared with the situation in which the abstraction was learned. *Ibid.*, p. 120.

^d Analysis is defined as "the breakdown of the material into its constituent parts and detection of the relationships of the parts and of the way they are organized. . . . Analysis . . . may be divided into three types or levels. At one level, the student is expected to break down the material into its constituent parts, to identify or classify the elements of the communication. At a second level, he is required to make explicit the relationships among the elements . . . their connections and interactions. A third level involves recognition of the organizational principles, the . . . structure of . . . the communication as a whole." *Ibid.*, pp. 144-145.

^e Synthesis is defined as "the putting together of elements and parts so as to form a whole, . . . combining them in such a way as to constitute a pattern or structure not clearly there before. Generally, this would involve a recombination of parts of previous experience with new material, reconstructed into a new and more or less well-integrated whole." *Ibid.*, p. 162.

^f Evaluation is defined as the making of conscious judgments about "the value, for some purpose, of ideas, works, solutions, methods, materials, and the like. . . . For purposes of classification, only those evaluations that are or can be made with distinct criteria in mind are considered." *Ibid.*, pp. 185-186.

1.00 KNOWLEDGE

"Knowledge," as defined in the taxonomy, includes those objectives and test situations that emphasize the remembering of facts and ideas. In a knowledge item, the student is required either to recall or recognize what he has learned—in exactly, or almost exactly, the way he originally learned it. The questions will usually be posed in a somewhat different form than in the textbook, but the testing process is chiefly one of checking on the completeness and accuracy of the student's memory for material learned.

Test items that measure student knowledge outnumber all others in most standardized achievement tests and most teacher-made tests. Gerberich² assembled 227 test exercises illustrating every type of test item in current use. His collection was far more varied than would be found in a random sampling of standardized or teacher-made tests. Yet a later study of all these items, in terms of the taxonomy, revealed that 51 percent were classifiable in the knowledge category.³ Undoubtedly, one of the chief reasons for the predominance of knowledge items is that they are comparatively easy to construct.

Knowledge is classified under three headings: (1.10) Knowledge of Specifics, (1.20) Knowledge of Ways and Means of dealing with specifics, and (1.30) Knowledge of the Universals and Abstractions in a field.

1.10 Knowledge of Specifics

Here we are concerned with the student's memory of specific items of information, which have meaning and value in themselves. They often represent basic elements that the student must know if he is to become acquainted with a field and solve problems within the field. "Knowledge of Specifics" is divided into two subcategories: (1.11) Knowledge of Terminology and (1.12) Knowledge of Specific Facts.

1.11 KNOWLEDGE OF TERMINOLOGY Each subject field utilizes a large number of terms and symbols that constitute the "shorthand" used in communication. A student is unable to think and work effectively in the field unless he can make use of the most essential of these verbal and nonverbal symbols.

Multiple-choice items can be used very effectively for questions on terminology. For example, students may be asked to identify the names of structures in actual animals. A frog may be dissected so that the heart is exposed; a numbered string may be attached to each of the various heart structures and the students asked to state the name and function of each structure so labeled. The structures to be identified may be listed beside the diagram, and the student asked to indicate his responses by placing the correct letter or number after each structure in the list. The element of guessing may be reduced by numbering more structures than are given in the list.

² Raymond Gerberich, *Specimen Objective Test Items* (New York: David McKay Company, Inc., 1956).

³ Julian C. Stanley and Dale L. Bolton, book review of the "Taxonomy of Educational Objectives," *Educational and Psychological Measurement*, vol. 17 (Winter 1957), pp. 632-633.

HEART OF FROG (VENTRAL VIEW)⁴

Structures:

a. Carotid artery	<u>5 or 6</u>
b. Left auricle	<u>2</u>
c. Post caval vein	<u>4</u>
d. Pulmocutaneous artery	<u>8</u>
e. Right auricle	<u>10</u>
f. Superior vena cava	<u>9</u>
g. Systemic artery	<u>7</u>
h. Truncus arteriosus	<u>11</u>
i. Ventricle	<u>3</u>

This type of exercise may also be used in physical science to identify parts of a machine, electrical circuit, and the like.

The more conventional type of matching exercise may also be used for testing knowledge of terminology. A list of definitions may be accompanied by a list of terms (longer than the list of definitions, to reduce the element of guessing). The list of terms should include some that are likely to be confused by the inadequately prepared student. The following example⁵ is illustrative:

- | | |
|-----------------------------------------------------------------------------------------------------------------------------------|-------------------|
| 187. The process in which electrons are gained in the outermost orbit of the atoms of the element. | A. Hydrogenation |
| 188. The process in which the hydrogen and hydroxyl ions in solution unite to form water and the other ions unite to form a salt. | B. Ionization |
| 189. The process in which electrons are lost from the outermost orbit of the atoms of the element. | C. Neutralization |
| | D. Oxidation |
| | E. Reduction |

Standardized tests in special subject fields include a number of items on the understanding of terminology, such as the following examples from the *Anderson Chemistry Test*.⁶

- (1) The valence of an element tells
 - (a) its atomic weight.
 - (b) the solubility of its compounds.
 - (c) its stability.
 - (d) how many electrons its atom lends, borrows, or shares.
 - (e) how many compounds can be formed.
- (2) Any solution which conducts an electric current is called
 - (a) an ion.
 - (d) an electrode.
 - (b) an electrolyte.
 - (e) a catalyst.
 - (c) a non-electrolyte.

⁴ Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann, *The Construction and Use of Achievement Examinations* (Boston: Houghton Mifflin Company, 1936), pp. 229-231.

⁵ Max D. Engelhart, "Evaluation of Achievement in Chemistry," *Journal of Chemical Education*, vol. 28 (July 1951), pp. 373-379.

⁶ *Anderson Chemistry Test*, Form Am, items 1, 5. Copyright 1950 by Harcourt, Brace & World, Inc., New York, N. Y. Copyright in Great Britain. All rights reserved. Reproduced by special permission.

The first item illustrates the technique of having the student choose the best *definition* for a given term in chemistry; in the second item, the definition is given, and the student is merely required to match it with the correct *term*. Items of the first type ordinarily require a higher level of understanding of the term or concept. Both types, however, require recognition only. For this reason, teachers may prefer to have students define essential terms in their own words, even though the scoring of such student definitions must be subjective.

1.12 KNOWLEDGE OF SPECIFIC FACTS Under this heading can be included test items that involve knowledge of important dates, events, persons, and places which help the student in thinking about specific problems or topics. Facts differ from terminology in that terminology usually represents the terms and symbols that authorities have agreed to use, whereas facts can be verified by means other than a consensus among workers in the field.

The term "specific" is used to differentiate facts that can be known as discrete elements from those that have meaning only in a larger context. In this sense, approximate information, such as the approximate time span covered by the Reconstruction Period, would be included. Knowledge about specific sources of information, such as the *Buros Yearbooks* for a student of measurement, would also be included in this category.

An attempt should be made in selecting published tests or in developing local tests, to measure knowledge of facts that are *relevant* to important concepts or principles. For example, the two questions below are relevant to the concept that there is strength in unity and that disunity weakens a people in its defense against its enemies.

1. One of the chief reasons for the fall of Greece was the fact that:
 - a. the Greeks were not good fighters
 - b. the city-states would not work and fight together
 - c. Greece was a country which could be easily invaded
 - d. the Greek cities had no forts or protecting walls
2. Which one of the following was a reason for the white man's success in taking territory from the Indians?

The Indians were:

- a. not experienced fighters
- b. often willing to fight on the side of the white men against other Indian tribes
- c. not able to ride or shoot as fast as the white men
- d. too few in number to fight the colonists successfully⁷

⁷ Reprinted with the permission of the California Test Bureau from Georgia Sachs Adams and John A. Sexson, *California Tests in Social and Related Sciences, Elementary, Part I, Form AA* (Monterey, Calif.: California Test Bureau, 1953), items 18, 50.

Negatively stated questions can often be used to advantage in testing knowledge of specific facts. In some areas, it is extremely difficult to devise three or four plausible distractors. It may be much easier and more satisfactory to construct some negatively stated questions that will challenge the students.

- (1) All of the following secrete hormones EXCEPT the
 - (a) pituitary gland
 - (b) lymph nodes
 - (c) parathyroid glands
 - (d) adrenal glands
 - (e) islets of the pancreas⁸
- (2) The only source of heat given below that may NOT produce carbon monoxide is
 - (a) gasoline
 - (b) electricity
 - (c) kerosene
 - (d) coal
 - (e) oil⁹
- (3) Four of the following were commonly included as qualifications for voting during the colonial period. Which one was NOT?
 - (a) Male sex
 - (b) Ownership of property
 - (c) Membership in the state church
 - (d) Ability to read and write
 - (e) Status of "free man"¹⁰

1.20 Knowledge of Ways and Means of Dealing with Specifics

In this subdivision of the taxonomy, we are concerned with the student's knowledge of the ways in which man has learned to systematize, study, and criticize facts and ideas. This is a fairly limited group of behaviors because we are focusing on the *knowledge* of these ways, rather than their actual application to specific problems. The objectives and test items under this heading are classifiable into five groups, as shown in Table 11.1.

1.21 KNOWLEDGE OF CONVENTIONS Under this subclass are included the usages, styles, and practices that are agreed upon or conventional. Examples would include the rules used in typing footnotes and bibliographies to insure consistency of style, conventional rules of social behavior, and accepted usages in spoken and written English.

One of the basic problems in testing knowledge of such conventions as

⁸ Nelson *Biology Test*, Form Bm, item 7. Copyright 1950 by Harcourt, Brace & World, Inc., New York, N. Y. Copyright in Great Britain. All rights reserved. Reproduced by special permission.

⁹ Read *General Science Test*, Form Bm, item 2. Copyright 1950 by Harcourt, Brace & World, Inc., New York, N. Y. Copyright in Great Britain. All rights reserved. Reproduced by special permission.

¹⁰ Reprinted with the permission of the Educational Testing Service from *Cooperative American Government Test*, Form X, item 22 (Princeton, N. J.: Educational Testing Service, 1947).

Social trends

- (1) Which statement best describes the change in the power of state and federal governments during the last twenty-five years?
- No important increase or decrease of power is noticeable.
 - Both state and federal powers have decreased.
 - State power has increased; federal power has decreased.
 - Both have increased in power, but the power of the states has increased more rapidly.
 - Both have increased in power, but the power of the federal government has increased more rapidly.¹⁴
- (2) On your answer sheet, mark the number of your answer
- if higher in 1950 than in 1900;
 - if lower in 1950 than in 1900.
- Percent of people over 65 years of age.
 - Death rate (deaths per 1000 population).
 - Divorce rate (divorces per 1000 population).
 - Percent of people living in cities.¹⁵

Cause-effect relationships

- (1) Three of the following were causes of the conflict between England and Spain in the New World. One was a result. Which was the result?
- growth of foreign trade of both England and Spain
 - supremacy of England on the sea
 - English slave trade with Spanish colonies
 - capture of Spanish galleons by Sir Francis Drake¹⁶
- (2) What conditions contributed to the economic depression of the early 1930s? Choose a, b, c, d, or e.
- The lack of farm prosperity in the 1920s.
 - The decline of foreign markets after World War I.
 - The lack of purchasing power of low-income groups.
 - The large military budgets of the 1920s.
 - The lack of industrial capacity and natural resources.
- 1, 2, 3
 - 1, 2, 4
 - 2, 3, 5
 - 1, 4, 5
 - all of the above¹⁷

¹⁴ Reprinted by permission of the Educational Testing Service from *Cooperative American Government Test*, Form X, item 58 (Princeton, N.J.: Educational Testing Service, 1947).

¹⁵ *Dimond-Pflieder Problems of Democracy Test*, Form Am, items 62-65. Copyright 1953 by Harcourt, Brace & World, Inc., New York, N. Y. Copyright in Great Britain. All rights reserved. Reproduced by special permission.

¹⁶ Reprinted with the permission of The California Test Bureau from *California Tests in Social and Related Sciences, Advanced*, Form AA, Part I, item 19 (Monterey, Calif.: California Test Bureau, 1954).

¹⁷ *Crary American History Test*, Form Am, item 88. Copyright 1950 by Harcourt, Brace & World, Inc., New York, N. Y. Copyright in Great Britain. All rights reserved. Reproduced by special permission.

1.23 KNOWLEDGE OF CLASSIFICATIONS AND CATEGORIES As a subject field becomes well developed, scholars develop classifications and categories that may seem arbitrary to the beginning student but that are fundamental to further work and study in the field.

Recall questions might be concerned with knowledge of the types of reliability coefficients, the subdivisions of the Dewey Decimal system, the headings of the taxonomy, the various classifications of jobs within the engineering profession or some other occupation, or the logical subdivisions within the biological sciences.

Test items that require the student to classify *unfamiliar* phenomena into categories would not fall in this subdivision. For example, an item testing the student's recall of the headings of the taxonomy would be classifiable here; whereas one requiring the classification of test items according to the taxonomy would fall under 3.00 Application, to be discussed later.

The following three examples illustrate the type of objective items that would be classifiable under this heading:

- (1) In all fairly complex animals the skeleton and the muscles are developed from the primary germ layer known as the
 1. ectoderm
 2. neurocoele
 3. epithelium
 4. endoderm
 5. mesoderm
- (2) Which of the following is a chemical change?
 1. Evaporation of alcohol
 2. Freezing of water
 3. Burning of oil
 4. Melting of wax
 5. Mixing of sand and sugar¹⁸
- (3) A triode radio tube differs from a diode tube in that it has:
 - (a) a lower vacuum
 - (b) a cathode
 - (c) a shield
 - (d) a grid
 - (e) an inert gas within the tube¹⁹

In each case, the student is required to recall the knowledge as taught, or to recognize a textbook interpretation, rather than to apply the system of categories in a new situation.

1.24 KNOWLEDGE OF CRITERIA This subclass is similar to the preceding one in that (1) it involves knowledge found useful by specialists in the field and (2) it includes only knowledge of criteria rather than their application to new problems.

¹⁸ Benjamin S. Bloom, ed., *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain* (New York: David McKay Company, Inc., 1956), p. 83.

¹⁹ *Dunning Physics Test*, Form Am, item 75. Copyright 1950 by Harcourt, Brace & World, Inc., New York, N. Y. Copyright in Great Britain. All rights reserved. Reproduced by special permission.

Since considerable emphasis has been placed in this textbook on knowledge of criteria for test selection, two illustrative items are given from this subject area.

1. What is characteristic of a highly reliable test?
 - (A) Two scores for the same examinee agree closely with each other.
 - (B) Good students make much higher scores on the test than poor students.
 - (C) There is a uniform distribution of scores on the test.
 - (D) The items in the test vary widely in difficulty.
2. In general, low reliability is associated with a
 - (A) relatively large error of measurement
 - (B) relatively small error of measurement
 - (C) below norm performance
 - (D) non-representative norms
 - (E) relatively high validity²⁰

1.25 KNOWLEDGE OF METHODOLOGY Knowledge concerning the procedures employed in a subject field is an important prerequisite to the use of such methodology in studying new problems. Before engaging in any type of inquiry, the student should know the techniques that have been effectively used for the study of similar problems.

Again examples from the measurement area are used.

- (1) In the scoring of essay examinations, all the following are generally considered desirable practices except to
 - (A) reduce the mark for poor spelling or penmanship
 - (B) prepare a scoring key and standards in advance
 - (C) remove or cover pupils' names from the papers
 - (D) score one question on all papers before going to the next
 - (E) use the same standards for all pupils²¹
- (2) Describe one method of making an item analysis which can be used to study both item difficulty and item discrimination.

1.30 Knowledge of the Universals and Abstractions in a Field

Ideally, a student's knowledge of a field is not limited to specifics and ways of dealing with them, but goes beyond this to a knowledge of the major ideas and patterns by which these facts and ideas are organized. A

²⁰ Reprinted with the permission of the authors, Robert L. Ebel and Eric F. Gardner, *Multiple-Choice Items for a Test of Teacher Competence in Educational Measurement* (Ames, Iowa: National Council on Measurement in Education, 1962), p. 23.

²¹ Reprinted by permission of the publisher from Victor H. Noll and Joe L. Saupe, *Manual for Introduction to Educational Measurement* (Boston: Houghton Mifflin Company, 1957).

student who can recall many abstractions in a subject field, and can recall specific illustrations of them that have been studied in class, has the basis for relating and organizing a great many specifics. As a result, he tends to gain greater insight into large units of subject matter and to show greater retentiveness for both the generalizations and the supporting facts.

In this category of objectives and test items, we have only two major subclassifications: (1.31) knowledge of principles and generalizations and (1.32) knowledge of theories and structures.

1.31 KNOWLEDGE OF PRINCIPLES AND GENERALIZATIONS Here we are concerned with abstractions that help us to describe, explain, or predict phenomena. The test items are ones that require (1) that the student know the principles and generalizations in the sense that he can recognize or recall correct versions of them, or (2) that the student recall or recognize specific illustrations of these generalizations that have been used in the textbook, or other instructional materials.

Examples:

- (1) When a gas is heated at constant volume, its pressure increases because
 - (A) its molecules increase in size although their speed remains constant
 - (B) the average change in momentum per molecule hitting the walls of the container increases
 - (C) Charles' law is true
 - (D) its molecules increase in both size and speed
 - (E) clusters of molecules break apart and fly about separately²²
- (2) The Constitution of the United States provided long terms for federal judges in order to
 - (A) render it impossible for them to serve in other positions in the federal government
 - (B) eliminate expenses that accrue with frequent elections
 - (C) facilitate the rendering of decisions without political influence
 - (D) make the federal bench an attractive career service
 - (E) secure continuity in the administration of justice²³

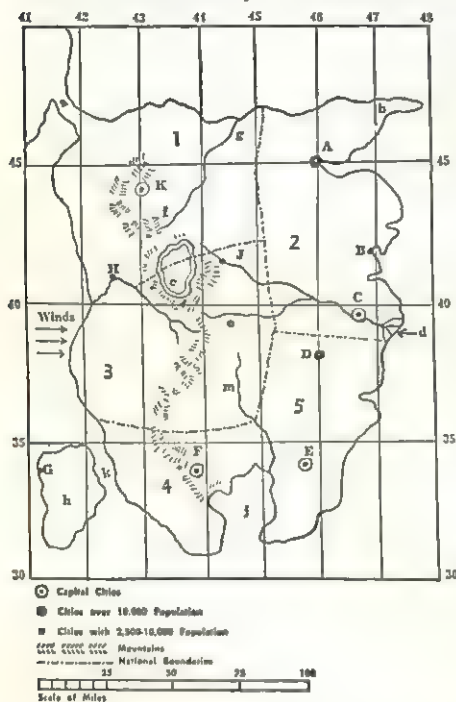
1.32 KNOWLEDGE OF THEORIES AND STRUCTURES This subcategory represents the highest level of abstraction of the entire Knowledge category. It differs from category 1.31 in that it involves knowledge of a body of interrelated principles and generalizations. In order to perform successfully on items in this classification, the student must have a clear and systematic overview of some theory, such as the theory of evolution, or of some structure, such as the structural organization of the city, state, or Federal government.

²² *A Description of the College Board Achievement Tests* (Princeton, N. J.: The College Entrance Examination Board, 1962), p. 142.

²³ *Ibid.*, p. 108.

completion type, which require students to fill in specified data read from a graph or table, such as the mileage between two towns, the name of a capital city, and the like.

✓ Below is given a map of a make-believe continent. There are five countries, numbered 1, 2, 3, 4, and 5. Read each question below, and then use the map to answer it. For each question, mark the answer as you have been told.



76. Which one of the following cities is the capital of Country 4?
D E F _____76
77. Which one of the following cities is the largest?
A B G _____77
78. Which city is slightly northeast of City H?
K B D _____78
79. Which one of the following cities is farthest from the equator?
H E F _____79
80. The distance from City C to City H is about
a 100 miles b 50 miles c 75 miles _____80
81. Between which two countries does a river form part of the boundary?
d 1 and 2 e 1 and 3 f 2 and 5 _____81

✓ The locations of certain physical features have been indicated on the map by small letters. Find each physical feature, then mark the answer as you have been told.

82. Lake b c h _____82
83. Delta b d j _____83
84. Isthmus a e k _____84
85. Source of stream f g m _____85

Test 3 - Sec. C Score
(number right) _____

Fig. 11.1 An Illustrative Test Item Involving Skill in Translation.

Reproduced with the permission of the California Test Bureau from Georgia Sachs Adams and John A. Sexson, *California Tests in Social and Related Sciences, Elementary, Form AA* (Monterey, Calif.: California Test Bureau, 1955).

A third type of translation is from one verbal form to another, as in translating a communication from a foreign language to English, or from a poetic or dramatic form (with its symbolism and metaphors) to everyday prose.

2.20 Interpretation and 2.30 Extrapolation

The essential behavior in Interpretation is the comprehension of the major ideas in a communication and an understanding of their interrelationships. Interpretation goes beyond translating *parts* of a table, map, or other communication to a determination of the larger, more general ideas that can be drawn from it. One of the questions on the map, no. 79, went beyond translation in that the student was asked to consider the relationship between the various parts, to get a view of the map as a whole, and relate it to his own fund of concepts. That is, "interpretation" was involved when we asked the student "Which one of the cities is farthest from the equator?"

In the following algebra problems, the student must comprehend not only the elements but their interrelationships in order to work them successfully.

- (1) The area of a certain square is represented by the expression $9x^2 + 6xy + y^2$. What will be the perimeter P of this square expressed in terms of x and y ?²⁷
- (2) If h , k , m , and n are positive numbers, k is greater than m , and n is greater than h , which of the following is (are) true?
 - I. $n + h$ may equal $k + m$.
 - II. $k + h$ may equal $n + m$.
 - III. $k + n$ may equal $m + h$.(A) None (B) I only (C) I and II only (D) I and III only (E) I, II, and III²⁸

Competency in extrapolation, or making inferences from trends or relationships in the data, requires a recognition that the inference involves some degree of probability. Extrapolation involves generalizing from a sample to a universe, from one situation to similar situations, from a trend in the past to a prediction for the future.

It is often efficient to test for competency in extrapolation by devising items on the same map, graph, or table that was used to test the interpretation objectives. If the exercise is an objective one, we can include some generalizations that involve a sound extension of evident trends and others that clearly involve overgeneralization.

In interpretation of data tests, the student can be presented with a prose selection, table, chart, or map and asked to supply or recognize inferences that can legitimately be made from the data. Wherever possible, these inferences should be based on the joint consideration of two or more elements in the communication.

²⁷ Hawkes, Lindquist, and Mann, *op. cit.*, p. 367.

²⁸ A Description of the College Board Achievement Tests, *op. cit.*, p. 131.

Directions: Below are some statistics relating to education and occupations. You are to judge what conclusions may be drawn from them.

	Occupational distribution found in a sample of male college graduates*	Distribution of occupations in the population as a whole, 1950
OCCUPATIONS	PERCENTAGES	
Executives, minor officials, partners, proprietors	23.5	9.1
Professional workers	51.3	4.7
Salesmen	6.0	Less than 1%
Skilled workers	7.1	33.8
Clerical workers	8.7	13.4
Unskilled workers	1.7	26.1
Farmers	1.7	13.0
	<hr/> 100.0	<hr/> 100.0

* You may assume that the sample selected is representative of all male college graduates in the United States.

Below are a series of statements relating to occupations and education.

Blacken answer space

- A—if the foregoing statistics alone are sufficient to prove the statement true;
 - B—if the foregoing statistics alone are sufficient to indicate that the statement is probably true;
 - C—if the foregoing statistics alone are not sufficient to indicate whether there is any degree of truth or falsity in the statement;
 - D—if the foregoing statistics alone are sufficient to indicate that the statement is probably false;
 - E—if the foregoing statistics alone are sufficient to prove the statement false.
30. Typically farmers are completely uneducated.
 31. The professions absorb a larger percentage of male college graduates than any other group in the country.
 32. Sons of unskilled workers and sons of farmers have an approximately equal chance to go to college.
 33. Educational opportunity for the lower classes is increasing.
 34. The same proportions of farmers and of unskilled workers are college graduates.²⁹

Items of this type should be used as test questions only when students have already been given class exercises in the use of these terms. Otherwise the student may use responses (B) and (D) to indicate his doubt about his own answer, rather than his doubt concerning the adequacy of supporting data.

By means of test exercises of this type, the teacher can obtain evidence concerning (1) the student's ability to recognize the truth or falsity of gen-

²⁹ Benjamin S. Bloom, *op. cit.*, p. 110.

eralizations that are clearly supported or negated by the data, (2) his ability to identify inferences for which the correct response is "insufficient data"; and (3) his tendency to "go beyond the data" or to be overcautious in extrapolating from the data to infer trends.³⁰

The following question illustrates how we can test the student's ability to extrapolate by making inferences regarding the probable point of view of a historical leader, a political party, or some other type of leader or group.

"What choices, then, are left us in the realm of foreign policy? I see only two: imperialistic adventuring and the active promotion of world peace, and which of these alternatives is likely to supply the more favorable conditions for the continuance of constitutional democracy among us is hardly open to reasonable doubt. Even wars fought for the most generous ends can still spell disaster for that complex set of values which our Constitution aims to uphold and promote."

The words most nearly reflect the sentiments of

- (A) George Washington
- (B) Abraham Lincoln
- (C) Woodrow Wilson
- (D) Theodore Roosevelt
- (E) Thomas Jefferson³¹

This question tests both knowledge and extrapolation. Extrapolation items in the interpretation-of-data exercises provide a purer measure of competency in extrapolation by providing the data from which inferences can be made.

3.00 APPLICATION

The chief difference between the categories of Comprehension and Application is that the latter involves facing a new problem, or a problem that appears quite unfamiliar until the student has restructured the elements into a familiar context. Comprehending an abstraction does not guarantee that the individual will be able to recognize its relevance and apply it correctly in real-life situations. Students need practice in restructuring unfa-

³⁰ By the use of a stencil scoring key, the position of each correct answer can be indicated on each student's answer sheet. Scores for "going beyond the data" or overgeneralizing can be obtained by counting incorrect answers that are in the direction of the extremes of the scale, while the "over caution" score is obtained by counting incorrect answers in the direction of the center of the scale. "Crude errors" are errors that traverse the center of the scale, that is, are on the opposite side of center from the keyed response.

³¹ *A Description of the College Board Achievement Tests, op. cit.*, p. 111.

miliar problem situations and applying the concepts and principles they have learned.

As our technological world has become more complex and more rapidly changing, the application of learnings to new problems has become even more important. Hence, the effectiveness of education cannot be adequately appraised unless we find out how well students can apply what they have learned in situations that differ from the textbook situations in which the concepts and principles were originally studied.

The test items presented here as representing the Application category might involve only knowledge or comprehension if the specific problem presented in the test item had been studied in the textbook or in classroom discussions. In order to test Application, a test situation must be new to the student or contain new elements that require rethinking or restructuring of the material learned. In our attempts to set up new situations, three approaches are useful: (1) presenting a fictional situation, (2) using material with which students are not likely to have had contact (such as simplified versions of complex problems studied in more advanced work), and (3) taking a new slant on common situations.³²

In some Application items, such as the following, the *process* by which the student reaches the solution of the proposed problem is not shown.

- (1) If the earth were viewed from the moon, which one of the following statements would be true? The earth would
- appear black because it generates no light.
 - show phases similar to those we see on the moon.
 - appear about the same size as the moon does to us.
 - eclipse the sun during a considerable part of each month.³³

In other items, practically the entire process of choosing the correct principles and applying them in the solution of the problem is recorded. The following essay item is an example:

John prepared an aquarium as follows. He carefully cleaned a ten-gallon glass tank with salt solution and put in a few inches of fine washed sand. He rooted several stalks of weed (*elodea*) taken from a pool and then filled the aquarium with tap water. After waiting a week, he stocked the aquarium with ten one-inch goldfish and three snails. The aquarium was then left in a corner of the room. After a month the water had not become foul and the plants and animals were in good condition. Without moving the aquarium he sealed a glass top on it.

What prediction, if any, can be made concerning the condition of the aquarium after a period of several months?

³² Benjamin S. Bloom, *op. cit.*, p. 130.

³³ Reprinted by permission of The California Test Bureau from Georgia Sachs Adams *et al.*, *California Tests in Social and Related Sciences*, Advanced, Form AA (Monterey, Calif.: California Test Bureau, 1954), Test 5, item 7.

If you believe a definite prediction can be made, make it and then give your reasons. If you are unable to make a prediction for any reason, indicate why you are unable to make a prediction (give your reasons).³⁴

In the following objective item, the entire process of selecting a correct or incorrect conclusion and supporting the conclusion by relevant or irrelevant reasons is clearly recorded.

An electric iron (110 volts, 1000 watts) has been used for some time and the plug contacts have become burned, thus introducing additional resistance. How will this affect the amount of heat which the iron produces?

Directions: Choose the conclusion which you believe is most consistent with the facts given above and most reasonable in the light of whatever knowledge you may have, and mark the appropriate space on the Answer Sheet.

Conclusions:

- A. The iron will produce more heat than when new.
- B. The iron will produce the same heat as when new.
- C. The iron will produce less heat than when new.

Directions: Choose the reasons you would use to explain or support your conclusion and fill in the appropriate spaces on your Answer Sheet. Be sure that your marks are in one column only—the same column in which you marked the conclusion.

Reasons:

- 1. The heat produced by an electrical device is always measured by its power rating. It is independent of any contact resistance.
- 2. Electric currents of the same voltage always produce the same amount of heat, and burned contacts do not decrease the amount of electricity entering the iron.
- 3. The current which flows through the iron is reduced when the resistance is increased.
- 4. Increasing the resistance in an electrical circuit increases the current.
- 5. An increase in electrical resistance increases the heat developed.
- 6. Manufacturers of electric irons urge that the contacts be kept clean to maintain maximum efficiency.
- 7. An increase in the temperature of a wire usually results in an increase in its resistance.
- 8. Burned contacts increase the heat developed in an electric iron just as increasing the friction in automobile brakes develops more heat.
- 9. The heat developed by an electric iron when connected to 110 volts is independent of the flow of current.³⁵

In most cases, we administer Application items in order to evaluate the effectiveness of instruction and the growth of students toward course objectives. Hence, we are not interested in the extent to which students can

³⁴ Adapted from PEA Test 1.3 B, "Application of Principles in Science," Evaluation in the Eight-Year Study, cited by *The Measurement of Understanding*, Forty-fifth Yearbook, Part I (Chicago: National Society for the Study of Education, 1946), p. 111.

³⁵ Problem VI from PEA Test 1.3, "Application of Principles," cited by Benjamin S. Bloom, *op. cit.*, p. 132.

solve a problem by common sense or on the basis of common knowledge; rather, we are interested in the extent to which the student has learned to apply the concepts and generalizations taught in a specific course. One must therefore guard against the inclusion of clues to the solution of a problem that would be helpful to the bright student without specialized knowledge. Perhaps the best safeguard against items that evaluate general problem-solving ability is to administer them to persons who equal our students in intelligence but have not taken the specific course for which the exercise was designed.

4.00 ANALYSIS

Skill in analysis is included as an objective in many subject fields. Teachers of science and social studies want their students to be able to distinguish facts from hypotheses in both written and spoken communications, distinguish major from subordinate ideas, and recognize when unstated assumptions are involved in reaching a conclusion. Teachers of music and literature want students to be able to distinguish dominant and subordinate themes, to find evidence of a composer's or author's techniques and purposes.

Analysis implies breaking down a communication into its parts, plus seeing the way in which the parts are organized in relationship to each other. At the lowest level of analysis, the student is expected to identify and classify the *elements* of the communication, for example, differentiating between hypotheses and conclusions, or recognizing the unstated assumptions being made by the author. At the second level, analysis is concerned with the study of *relationships* among the elements of a communication, or among the various parts of a document; for example, the relevance of supporting points to a central idea, or the relationship of a hypothesis to evidence presented in support of it. The third and highest level of analysis involves the analysis of *organizational principles*. An author or composer rarely points out the organizational principles he has used in developing a speech, a poem, a play, or a symphony. Yet the reader or listener may not achieve full understanding of a communication until he has discerned a speaker's underlying point of view, identified the techniques the artist is using, or recognized in a play or sonnet its underlying structure or pattern.

Test items in this field should present the student with new material, rather than material that has already been analyzed in the text or in class discussions. The material presented for analysis may be an unfamiliar selection from literature, a report of a new experiment, a description of a hypothetical social situation, or an unfamiliar picture or musical composition. The test items may be of either the essay or objective type. When objective-

type exercises are to be constructed, it is best to first administer parallel essay items and analyze student responses to them. As a result of such analysis the distractors or wrong alternatives in the objective test can represent errors in analysis that students actually make.

In a standardized test of critical thinking, the following exercise on "Inference" is included.³⁶

Directions:

- T if you think the inference is definitely TRUE; that it properly follows beyond a reasonable doubt from the statement of facts given.
- PT if, in the light of the facts given, you think the inference is PROBABLY TRUE; that there is better than an even chance that it is true.
- ID if you decide that there are INSUFFICIENT DATA, that you cannot tell from the facts given whether the inference is likely to be true or false; if the facts provide no basis for judging one way or the other.
- PF if, in the light of the facts given, you think the inference is PROBABLY FALSE; that there is better than an even chance that it is false.
- F if you think the inference is definitely FALSE; that it is wrong, either because it misinterprets the facts given, or because it contradicts the facts or necessary inferences from those facts.

In 1946 the United States Armed Forces conducted an experiment called "Operation Snowdrop" to find out what kinds of military men seemed to function best under severe Arctic climatic conditions. Some of the men selected came from Northern European stock while others came from Latin or Mediterranean stock; some were stout and some were thin; some were draftees and some volunteers; some had normal blood pressure while some had slightly high or low blood pressure. All of the participants in "Operation Snowdrop" were given a training course in how to survive and function in extreme cold. At the conclusion of the experiment it was found that the only two factors among those studied which distinguished between men whose observable performance was rated as "effective" and those rated as "ineffective" on the Arctic maneuvers were: (1) desire to go (volunteer versus draftee), and (2) degree of knowledge and skill regarding how to live and protect oneself under Arctic conditions.

- 11. Despite the training course given to all of the participants in "Operation Snowdrop," some exhibited greater Arctic survival knowledge or skill than others.....
- 12. The Armed Forces expected that important future military operations might be carried on in the Arctic.....
- 13. A majority of the men who participated in "Operation Snowdrop" thoroughly disliked that experience.....
- 14. As a group, the men of Nordic backgrounds were found able to withstand the cold and to function more effectively than those of Latin backgrounds.....
- 15. Participants who were normal in weight and blood pressure were found much better than other participants at acquiring skills to protect themselves under Arctic conditions.....

³⁶ *Watson-Glaser Critical Thinking Appraisal*, Revised Form Zm, items 11-15. Copyright 1961 by Harcourt, Brace & World, Inc., New York, N. Y. Copyright in Great Britain. All rights reserved. Reproduced by special permission.

Test items on analysis of relationships are especially suitable for open-book tests. The student can show his ability to check the relevance of points made by the author to his central idea.

The following item is based on an excerpt from Lindsay's *The Modern Democratic State*.³⁷

The relation between the definition of *sovereignty* given in Paragraph 2 and that given in Paragraph 9 is best expressed as follows:

1. There is no fundamental difference between them, only a difference in formulation.
2. The definition given in Paragraph 2 includes that given in Paragraph 9, but in addition includes situations which are excluded by that given in Paragraph 9.
3. The definition given in Paragraph 9 includes that given in Paragraph 2, but in addition includes situations which are excluded by that given in Paragraph 2.
4. The two definitions are incompatible with each other; the conditions of sovereignty implied in each exclude the other.³⁸

The following items are based on a musical composition, which is played for the students.

- (1) The general structure of the composition is
 1. theme and variations.
 2. theme, development, restatement.
 3. theme 1, development; theme 2, development.
 4. introduction, theme, development.
- (2) The theme is carried essentially by
 1. the strings.
 2. the woodwinds.
 3. the horns.
 4. all in turn.³⁹

5.00 SYNTHESIS

Synthesis involves the student's combining elements or parts in such a way as to constitute a pattern or structure that is new to him. As a rule, new experiences or materials are combined with those previously learned into a new integration. This category is the one that most clearly provides for creative activity on the part of students.

Not all essay questions belong under Synthesis. In many essay questions, the student is required only to translate what he recalls of the textbook statement into his own words. Or an essay question may simply require the

³⁷ Alexander D. Lindsay, *The Modern Democratic State* (New York: Oxford University Press, 1947).

³⁸ Benjamin S. Bloom, *op. cit.*, p. 159.

³⁹ *Ibid.*, p. 161.

student to analyze the structural elements in a sonnet, or to recognize the structure of a musical composition.

Three subcategories under Synthesis are distinguished on the basis of the *products* of the creative restructuring process: (1) production of a unique communication, (2) production of a plan, or proposed set of operations, and (3) derivation of a set of abstract relations. It is assumed that these three fairly distinct kinds of products require somewhat different cognitive processes.

In a test item or assignment requiring synthesis the student is allowed considerable latitude with respect to the content of his communication and hence can draw freely upon his own ideas, feelings, and experiences. Yet he is not allowed completely free expression because the task is structured to show how much the student has grown with respect to such synthesis objectives as

Ability to make an extemporaneous speech

Ability to write an informative essay

Ability to write a short story (or a poem) that others would find interesting and entertaining

Ability to set a short poem to music.

When the Portland, Oregon, schools decided to select their most talented students in various fields, they developed five exercises in creative writing that could be classified as "unique communications." Two of the five exercises are briefly described.

Developing Expressive Sentences

After the children had had some preparatory work in this type of exercise, the children were given several sentences, for example, "The man went down the street." The children were asked: "In what way could you add to or change the word 'man' to give a clearer picture of the man? In what ways could you change other words in the sentence to make us see this man going down the street?"

Developing a Paragraph from a Sentence

After the children had had some preliminary experience with similar assignments, they were asked to choose one sentence from a group of suggested sentences and write a paragraph about it. Sentences which would stimulate the imagination of children were used, such as: "The mysterious box drew all eyes to it."⁴⁰

⁴⁰ Adapted from Robert C. Wilson, "Improving Criteria for Complex Mental Processes," *Invitational Conference on Testing Problems* (Princeton, N. J.: Educational Testing Service, 1957), pp. 14-17.

The following two sets of directions vary with respect to the amount of structuring of the exercise. The first permits much more freedom in choice of content than does the second.

- (1) "Think of some time in your own life when you were up against a difficulty, something that stood in your way and had to be overcome. Make up a story around this difficulty and tell it to the class."
- (2) "Think of a plot based upon an obstacle that could occur between the following two sentences, and then develop a short story using these sentences and your plot."
It was an event to be honored with a party, preferably a surprise party . . . "It was a surprise, all right—a surprise all the way around!"⁴¹

These examples have all been concerned with creative writing or extemporaneous speaking. However, similar test situations can easily be devised in art, music, creative dance, or creative dramatics.

In an exercise involving the development of a "plan or proposed set of operations," the requirements for the student's product are usually presented in the form of specifications to be met or data to be considered. The student is encouraged to develop his own approach to the problem. For example, a science student might be asked to propose ways of testing specific hypotheses. A student in teacher education might be asked to plan a unit of instruction to achieve specified objectives at a specified grade level; however, he would be encouraged to use his own ideas concerning specific content and activities.

In a class in sociology, students could be presented with relevant data concerning each of two or more communities within a metropolitan area. They could be asked to propose a plan for the reduction of juvenile delinquency in each area that would be consistent with the data given about various community factors usually associated with juvenile delinquency.⁴² Although such an exercise involves application of principles learned in sociology, the amount of latitude allowed the student in selecting and organizing his learnings makes such an exercise classifiable under synthesis.

Any exercise (for example, in such courses as industrial arts, homemaking, or commercial art) that requires students to design a product to meet a set of specifications or to satisfy a certain purpose would probably fall under this category. If such an assignment permitted the student to show individuality in his solution and independence in thought and action rather than dependence on textbook or teacher, abilities in synthesis would probably be involved.

The student who observes many related phenomena and formulates an

⁴¹ Benjamin S. Bloom, *op. cit.*, p. 178.

⁴² *Ibid.*, pp. 180–181.

hypothesis that adequately accounts for them is performing the third type of synthesis task. For example, the pupil who formulates *for himself* statements about the relationships between corresponding parts of similar triangles or makes similar discoveries in mathematics or science would be making progress toward a synthesis type of objective.

In testing for progress toward synthesis objectives, we need to help students feel free from pressure to conform to the views or preferred methods of the teacher. Too much control and too many instructions will stifle students' creativity. Enough time must also be allowed for the student to become thoroughly acquainted with an unfamiliar task, to explore possible approaches, and to reach a synthesis that seems best for him. Sometimes the time problem can be partially met by allowing the students to do some of their preparation before the examination period. Special reading materials that would help in the writing of an essay, for example, can be distributed and studied ahead of time.

The evaluation of the products of synthesis present another problem in that objective criteria of worth are often lacking. The independent judgments of qualified persons are perhaps the only basis for evaluation of many of these products. The problem of improving the validity and reliability of such subjective judgments is considered further in Chapter 12.

6.00 EVALUATION

Any person makes innumerable evaluations daily as he judges persons, objects, or activities as being more or less useful to him, more or less attractive, or as either enhancing or threatening to his status of self-esteem. Many of these evaluations are highly egocentric and quickly made without careful consideration. In the taxonomy, these are considered to be *opinions*, while the term "evaluation" is reserved for those evaluative judgments that are, or can be, consciously made with distinct criteria in mind. They usually require fairly adequate comprehension and analysis as a basis for judgment. The student can readily see how an evaluation of a standardized test in terms of the criteria presented in this textbook would differ from an opinion that the XYZ test, used by a highly regarded school district, would be a good one for local use.

Evaluation is placed last among the categories because it is a complex process that involves some combination of all the other behaviors. It can be defined as the making of judgments about the value, for some purpose, of an idea, method, solution, or product. The criteria used in making value judgments in a test exercise may either be given to the student or determined by him.

- (B) if the line is inappropriate in *rhythm or meter*,
 (C) if the line is inappropriate in *style or tone*,
 (D) if the line is inappropriate in *meaning*.

Hail, bards triumphant! born in happier days,
 Immortal heirs of universal praise!
 Whose honors with increase of ages grow

56. Like saplings stuck in dirt long years ago.
 57. As streams roll down, enlarging as they flow.
 58. As streams in desert lands where hot winds blow.
 59. From obscurity to fame the world around doth know.⁴⁴

The following essay question about a poem represents a type of item that could easily be adapted to the evaluation of many different products.

Write an essay of from 250 to 500 words, describing and evaluating the foregoing poem. In your description you should employ such terms as will reveal your recognition of formal characteristics of the poem. Your principles of evaluation should be made clear—although they should not be elaborately described or defended.⁴⁵

Essay items give the student an opportunity to demonstrate his competence, but they do not focus sharply on the desired behaviors. Bloom deplores the shortage of good objective items in the evaluation category. He concludes the taxonomy with this final statement: "Perhaps the greatest value of the taxonomy . . . is in pointing to the need for further study and development of testing techniques for measuring competence in evaluating documents, materials, and works."⁴⁶

SUMMARY STATEMENT

Unless teachers make a conscious effort to test such cognitive abilities as comprehension and application, their examinations tend to test only students' memory of specific facts and rules. Since many students limit their studying to the types of learnings that they believe the teacher will include in the next examination, this overemphasis on memory items can have a very deleterious effect on student achievement. Moreover, if test scores are to constitute an adequate basis for making inferences regarding student achievement in a course, all course objectives should be represented in the table of specifications for the test and should be given the appropriate emphasis in the design and selection of test items. Conscientious use of the taxonomy, which provides a basis for classifying both objectives and test items, can aid materially in increasing the content validity of teacher-made tests, as well as standardized ones.

⁴⁴ *A Description of the College Board Achievement Tests*, *op. cit.*, p. 36.

⁴⁵ Benjamin S. Bloom, *op. cit.*, p. 198.

⁴⁶ *Ibid.*, p. 195.

The concepts involved in the taxonomy are complex and interrelated ones. In order to help students to understand and use the taxonomy, the author has included in Table 11.1 illustrative objectives for each of the 20 subcategories. In the text of this chapter, an attempt has been made to define and illustrate each subcategory and to clarify distinctions between them. In providing test items illustrative of each subcategory, the author has attempted to include items that are desirable models of item writing and ones that represent a large number of subject areas. In this way, the specimen items can serve the dual purpose of illustrating the classifications of the taxonomy and of providing additional samples of item writing in mathematics, science, history, foreign language, and other subject areas.

SELECTED REFERENCES

- BLOOM, BENJAMIN S., ed., *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain*. New York: David McKay Company, Inc., 1956.
- COOK, DESMOND L., "The Use of Free Response Data in Writing Choice-Type Items," *Journal of Experimental Education*, vol. 27 (December 1958), pp. 125-133.
- CURETON, EDWARD E., "The Rearrangement Test," *Educational and Psychological Measurement*, vol. 20 (Spring 1960), pp. 31-35.
- EBEL, ROBERT L., "Writing the Test Item," in E. F. Lindquist, ed., *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, Chapter 7.
- ENGELHART, MAX D., "Suggestions for Writing Achievement Exercises to be Used in Tests Scored on the Electric Scoring Machine," *Educational and Psychological Measurement*, vol. 7 (Autumn 1947), pp. 357-374.
- GERBERICH, J. RAYMOND, *Specimen Objective Test Items: A Guide to Achievement Test Construction*. New York: David McKay Company, Inc., 1956.
- GRAHAM, G., *Teachers Can Construct Better Achievement Tests*. Curriculum Bulletin No. 170. Eugene, Ore.: University of Oregon, 1956.
- HAWKES, HERBERT E., E. F. LINDQUIST, AND C. R. MANN, *The Construction and Use of Achievement Examinations*. Boston: Houghton Mifflin Company, 1936, Chapter 7.
- The Measurement of Understanding*, 45th Yearbook, National Society for the Study of Education, Part I. Chicago: University of Chicago Press, 1946.
- NEDELSKY, LEO, "Ability to Avoid Gross Error as a Measure of Achievement," *Educational and Psychological Measurement*, vol. 14 (Autumn 1954), pp. 459-472.
- REINER, WILLIAM B., "Evaluating Ability to Recognize Degrees of Cause and Effect Relationships," *Science Education*, vol. 34 (February 1950), pp. 15-28.
- SMITH, E. R., R. W. TYLER, AND OTHERS, *Appraising and Recording Student Progress*. New York: Harper & Row, Publishers, Inc., 1942.
- WEITZMAN, ELLIS, AND WALTER J. MCNAMARA, "Apt Use of the Inept Choice in Multiple Choice Testings," *Journal of Educational Research*, vol. 39 (March 1946), pp. 517-522.
- WOOD, DOROTHY ADKINS, *Test Construction: Development and Interpretation of Achievement Tests*. Columbus, O.: Charles E. Merrill Books, Inc., 1960.

English and Speech

- BRANDENBURG, ERNEST, AND PHILIP A. NEAL, "Graphic Techniques for Evaluating Discussion and Conference Procedures," *Quarterly Journal of Speech*, vol. 39 (April 1953), pp. 201-208.
- CROWELL, LAURA, "Rating Scales as Diagnostic Instruments in Discussion," *Speech Teacher*, vol. 2 (January 1953), pp. 26-32.
- DIEDERICH, PAUL B., "Making and Using Tests," *English Journal*, vol. 44 (March 1955), pp. 135-140, 151.
- , "Self-Correcting Homework in English," in Helen Huus, ed., *Education: Intellectual, Moral, Physical*. Philadelphia: University of Pennsylvania Press, 1960, pp. 258-271.
- , "Testing in the New English Program," *English Record*, vol. 3 (Spring 1953), pp. 11-17.
- DRESSEL, PAUL L., AND L. B. MAYHEW, *Handbook for Theme Analysis*. Dubuque, Iowa: William C. Brown Co., 1954.
- GATES, ARTHUR I., *A List of Spelling Difficulties in 3,876 Words*. New York: Bureau of Publications, Teachers College, Columbia University, 1937.
- HARRIS, CHESTER W., "Measurement of Comprehension of Literature," *School Review*, vol. 56 (May, June 1948), pp. 280-289, 332-342.
- HUDDLESTON, EDITH, "Measurement of Writing Ability at the College-Entrance Level: Objective vs. Subjective Testing Techniques," *Journal of Experimental Education*, vol. 22 (March 1954), pp. 165-213.
- PALMER, OSMOND E., "Evaluation of Communication Skills," in Paul L. Dressel and Associates, *Evaluation in Higher Education*. Boston: Houghton Mifflin Company, 1961, pp. 192-226.
- SMITH, DORA V., ed., *The English Language Arts*. Commission on the English Curriculum, National Council of Teachers of English, Curriculum Series, vol. I. New York: Appleton-Century-Crofts, 1952, Chapter 18.
- SWERINGEN, MILDRED E., "Evaluation in the Language Arts Program," *Children and the Language Arts*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1955, Chapter 20.
- THOMAS, EDNA S., *Evaluating Student Themes*. Madison, Wisc.: University of Wisconsin Press, 1955.
- THOMAS, MACKLIN, "Construction Shift Exercises in Objective Form," *Educational and Psychological Measurement*, vol. 16 (Summer 1956), pp. 181-186.
- VORDENBERG, WESLEY, "How Valid Are Objective English Tests?" *English Journal*, vol. 41 (October 1952), pp. 428-429.

Foreign Language

- AGARD, FREDERICK B., AND HAROLD B. DUNKEL, *An Investigation of Second-Language Teaching*. Boston: Ginn and Company, 1948.
- CORNELIUS, EDWIN T., JR., *Language Teaching: A Guide for Teachers of Foreign Languages*. New York: Thomas Y. Crowell Company, 1954.
- MANUEL, HERSCHEL T., "The Use of Parallel Tests in the Study of Foreign Language Teaching," *Educational and Psychological Measurement*, vol. 13 (Autumn 1953), pp. 431-436.
- PIMSLEUR, P., "French Speaking Proficiency Test," *French Review*, vol. 34 (April 1961), pp. 470-479.

- PIMSLEUR, P., AND OTHERS, "Foreign Language Learning Ability," *Journal of Educational Psychology*, vol. 53 (February 1962), pp. 15-26.
- RAYMOND, JOSEPH, "A Controlled Association Exercise in Spanish," *Modern Language Journal*, vol. 35 (April 1951), pp. 281-291.
- SADNAVITCH, J. M., AND W. L. POPHAM, "Measurement of Spanish Achievement in the Elementary School," *Modern Language Journal*, vol. 45 (November 1961), pp. 297-299.
- SCHENK, ETHEL A., *Studies of Testing and Teaching in Modern Foreign Languages*. Madison, Wisc.: Dembar Publications, 1952.

Mathematics

- BROWN, CLAUDE H., *The Teaching of Secondary Mathematics*. New York: Harper & Row, Publishers, Inc., 1953, Chapter II.
- CLARK, JOHN R., ed., *Emerging Practices in Mathematics Education*. 22d Yearbook, National Council of Teachers of Mathematics. Washington, D.C.: The Council, 1954, Part 5.
- DONOVAN, JOHNSON, WITH OTHERS, "The Evaluation of Mathematical Learning," *Emerging Practices in Mathematics Education*. 22d Yearbook, National Council of Teachers of Mathematics. Washington, D.C.: The Council, 1950, Part 5, pp. 339-409.
- FAWCETT, HAROLD P., ed., *The Nature of Proof*. 13th Yearbook, National Council of Teachers of Mathematics. Washington, D.C.: The Council, 1941.
- MYERS, SHELDON S., *Published Evaluation Materials in Mathematics*. Annotated Bibliography of Mathematics Tests. Princeton, N.J.: Educational Testing Service, 1961. Reprinted from *Evaluation in Mathematics*. Washington, D.C.: The National Council of Teachers of Mathematics, 1961.
- PICKETT, HALE, *An Analysis of Proofs and Solutions of Exercises on Plane Geometry Tests*, Contributions to Education, No. 747. New York: Bureau of Publications, Teachers College, Columbia University, 1938.
- REEVE, W. D., "Evaluation Program in Secondary Mathematics," *School Science and Mathematics*, vol. 55 (February, March 1955), pp. 123-140, 216-228.
- SIMPSON, R. H., "Mathematics Teachers and Self-Evaluation Procedures," *Mathematics Teacher*, vol. 56 (April 1963), pp. 238-244.
- SPACHE, GEORGE, "A Test of Abilities in Arithmetic Reasoning," *Elementary School Journal*, vol. 47 (April 1947), pp. 442-445.

Reading

- BLOMMERS, PAUL, AND E. F. LINDQUIST, "Rate of Comprehension of Reading; Its Measurement and Its Relation to Comprehension," *Journal of Educational Psychology*, vol. 35 (November 1944), pp. 449-473.
- BRYAN, MIRIAM, "Can We Really Measure Reading Comprehension?—A Testing View," *The Journal of the Reading Specialist*, vol. 2 (September 1962), pp. 4-5.
- HUSBANDS, K. L., AND J. HARLAN SHORES, "Measurement of Reading for Problem Solving: A Critical Review of the Literature," *Journal of Educational Research*, vol. 43 (February 1950), pp. 453-465.
- NASLAND, R. A., AND OTHERS, "Evaluation and the Reading Program," *Claremont Colleges Reading Conference Yearbook*, 1961, pp. 133-141.

5. If tests of ability to apply principles were given annually to high school students of science, what influence could be expected on the instructional program in that area?

6. Prepare an exercise in which a student is asked to suggest or criticize procedures for testing an hypothesis in some area of science.

7. Select from standardized science tests a number of exercises that test science understandings and the ability to use the scientific method, rather than the memorization of isolated facts.

8. Members of a high school science faculty met to review a list of science objectives and to plan their evaluation program in terms of these objectives. One teacher seemed to express the feeling of the group when he said that the job seemed overwhelming. What practicable plans can you suggest to this group of teachers?

Evaluating Student Performance in the Skills

Even a casual comparison of a sampling of available tests with a typical list of educational goals would convince the reader that the evaluation of student performance in the skills has been grossly neglected. Almost all subjects have a number of important skills outcomes, such as laboratory skills in the sciences and the skills of handwriting, speaking, and effective writing in English instruction. Moreover, in fine arts, industrial arts, home-making, physical education, and vocational arts, the student's performance in various skills may assume even greater importance than his knowledge outcomes.

Because knowledges can be easily and efficiently measured by paper-and-pencil tests, teachers have rationalized their inattention to skills outcomes by assuming that there is a high relationship between knowledge and performance. Even examinations for admission to the bar, to teaching, and to medical practice have been largely verbal tests of examinees' knowledge of facts and principles.

Knowledge is necessary, but not sufficient, to adequate performance in the skills. Knowledge of traffic rules, although important, is no guarantee of ability to drive; knowledge of rules in athletics does not correlate highly with performance, nor knowledge of recipes and nutrition with ability to cook. "From the standpoint of validity one of the most serious errors committed in the field of human measurement has been that which assumes the high correlation of knowledge of facts and principles on the one hand and performance on the other."¹

Very little has been done to measure *performance in process* and the

¹ David G. Ryans and Norman Frederiksen, "Performance Tests of Educational Achievement," in E. F. Lindquist, *Educational Measurement* (Washington, D.C.: American Council on Education), p. 455.

products of performance, since (1) in such measurement it is difficult to obtain an adequate sampling of skills, and (2) it is difficult to evaluate attainment in many skills with even fair objectivity. However, the responsibility to measure student attainment is inescapable. If course objectives include the development of skills, other than those of a verbal nature, the use of performance tests and/or fairly objective ratings of products and processes are essential to effective teaching and learning.

In Chapter 5, tests were classified on the basis of the degree to which they directly measured criterion behavior. The most direct type was the work sample or "identical elements" test; the next type of test, involving some indirectness of measurement, was the "related behavior type." Most performance achievement tests are classifiable under one of these two types, that is,

1. The work sample type, in which the examinee is given a special opportunity under standard conditions to do some of the tasks on which we want to appraise his competency, such as sewing, cooking, or driving a car.
2. The simulated situation type (classifiable under "related behavior"), in which the examinee works in a test situation specially designed to be similar to the usual situation and to elicit the kinds of behavior we wish to measure (for example, students in a sewing class cut out the various pieces of a miniature dress pattern and pin them to a sheet of colored paper which represents dress goods).

Both these types of tests might be further classified into (a) those in which objective scoring is possible because there is a clear-cut distinction between rightness and wrongness (as in typewriting or mechanical assembly) and (b) those in which the scoring must depend on the judgment of the observer (as in instrumental or vocal performance, automobile driving, and the like).

DEVELOPING TESTS OF SKILLS OUTCOMES

As the teacher faces the task of evaluating student growth in manipulative and other physical skills, he recognizes the low validity of paper-and-pencil tests and the subjectivity of his daily observations. He may, therefore, consider the mastery of certain skills to be so important that they justify the development and use of performance tests. Such tests are especially valuable as a basis for diagnosis and reteaching. The results aid the instructor in assessing his own effectiveness in demonstrating certain skills; they reveal to him the points upon which greater emphasis should be placed.

Before we can score a student's performance or product, we have to select the specimens of behavior to be evaluated and plan the standard conditions under which they are to be obtained. Our procedures in setting up performance achievement tests will be similar to those described in Table 4.2 on content validity. That is, our basic approach will be (1) to define the universe of skills to be sampled and (2) to sample that universe by procedures that can be clearly described. As in other types of achievement tests, professional judgment will usually be required in the sampling process because it is seldom possible or efficient to use a random sampling of all the skills in a course or unit.

The universe of skills to be sampled is usually defined in a list of objectives for the course (which indicates the skills in which students should become proficient). The validity of the test will depend largely on the tasks selected to represent these general skills. If the sampling of skills to be included in the test is well done, coaching for the test should improve the student in the general abilities tested.

The following guidelines for selecting tasks for a skills test should improve test validity and the efficiency of measurement (per unit of testing time).

1. Choose tasks that are representative of the significant skills emphasized in the course.
2. Choose tasks, or aspects of tasks, that are reasonably difficult for students. Since performance testing is time consuming, eliminate tasks, or parts of tasks, that almost everyone can do. (For example, in a performance test of cooking skills, a student would fry bacon rather than potatoes; in a test of driving skills, a student would do parallel rather than angle parking.)
3. Plan a test that involves a minimum of repetition of identical procedures (for example, a test of driving skills should be planned so that the student spends little time repeating right turns and other routine tasks).
4. Choose tasks that are crucial to success on the job as a whole. The test should provide opportunities to make those mistakes that are frequently responsible for failure in the total task or for failure to progress to higher levels of proficiency. For example, a bandmaster would want to test students for accurate music reading and ability to come in at the proper place after a rest; a swimming teacher would place considerable weight on the student's ability to synchronize his breathing with the pattern of his crawl stroke, as well as those characteristics of swimming form that minimize water resistance.
5. If feasible, choose tasks that do not require too much time to perform, so that one can include a larger sampling of different tasks. (For example, a student of statistics might complete several problems in which the easier, time-consuming work had been done for him; or a student of instrumental music might play selected passages from several different compositions.)
6. Choose tasks in which the conditions of work can be made standard for all students and the performance can be judged with considerable objectivity. For example, the student's skill in frying eggs (see Figure 12.3) could probably be judged more objectively than his skill in frying potatoes. The student's skill in backing a car down a marked lane can be judged more

objectively than his ability in parallel parking. In the latter task, moreover, the observer must make allowance for the fact that cars differ in length and in ease of manipulation.

7. If possible, choose tasks that involve materials and equipment commonly used in the course, and of which there are sufficient sets available to permit a number of students to be tested at one time.

In any performance test, it is important to standardize conditions of work. Students must work under similar conditions if comparison is to be possible. In comparing students with respect to physical education skills, the position in the court from which the basketball is thrown and other specifications must be clearly indicated. In comparing shorthand students with respect to their skill in taking dictation, recorded dictation materials at specified dictation speeds should be used. When students' speeches or essays are compared, they should speak or write under similar conditions with respect to previous preparation, time allowed, type of subject assigned,² and the like. The teacher who grades students in their skills solely on the basis of products made at home may be comparing the products of a student who has received no help with those of another who has received the aid of a proficient parent and has been able to use special tools not available to all.

The following set of directions for a test of skill in laboratory techniques establishes standard working conditions and minimizes questions from students.

On the table you will find a supply of frogs and the materials needed to make a dissection.

1. Remove the skin from the frog's hind leg.
2. Dissect the gastrocnemius and the tibialis anticus longus muscles of the lower leg free from other muscles but left attached to the bones, showing their origin and insertion.
3. As soon as you have finished, notify the laboratory instructor so that the condition of the dissection can be scored before it has deteriorated.

You will be allowed exactly twenty minutes to make this dissection.³

Performance tests involve certain intrinsic difficulties: (1) the amount of time required in administration, since only a few students can usually be tested at a time; (2) the need for planning constructive activities for students not being tested at any given time; (3) the fact that such tests are

² Since the student's performance tends to vary with his interest and experience on an assigned topic, several samples on a variety of topics should be obtained.

³ Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann, *The Construction and Use of Achievement Examinations* (Boston: Houghton Mifflin Company, 1936), pp. 253-254.

difficult to construct and to administer; and (4) the fact that such tests tend to penalize students who cannot work well under pressure.

A test of laboratory techniques in college physics, in which 18 students could be tested at the same time, was developed by Kruglak. He set up 18 performance items, or stations, with a total of 35 possible responses. The verbal description of the given apparatus and the problem is typed on a 4- by 6-inch card, with each card being taped to the laboratory table next to the apparatus involved. At the beginning of the period, students are assigned to their initial stations at random. At three-minute intervals, as signaled by the teacher, each student moves to successive stations. Another test, designed to measure student competency in more complex techniques, involves six stations to which students are assigned for nine minutes each.⁴

SCORING PROCESSES AND PRODUCTS

Performance in process may be scored in terms of speed, use of approved methods, or general quality of the performance. Speed (of running, swimming, typewriting, writing of shorthand characters, and many other skills) is important and can be objectively measured. Notations with respect to the use of approved methods, are of considerable value in diagnosis and reteaching. The accuracy and quality of a process are usually judged in terms of their effect on the product. When this is impossible, as in vocal and instrumental music or physical skills, the subjective judgment of competent observers is used.

Relative Advantages of Scoring Processes and Products

The scoring of student performance in work samples or simulated-situation tests may be based on (1) *the performance in process*, (2) *the product* of the performance, or both. In some cases, the product is not distinguishable from the process, as in instrumental music, speech, or oral work in foreign language. In other cases, one can make a distinction, such as between the process of driving and the result (destination reached or distance traversed). In the case of driving, however, the process is all important; and subjective judgment concerning the process, made by competent observers, is indispensable. Fortunately, in many cases, the product

⁴ H. Kruglak, "Experimental Outcomes of Laboratory Instruction in Elementary College Physics," *American Journal of Physics*, vol. 20 (January 1952), pp. 138-139.

is most important; in the composing of music, for example, analysis of the process is less important than evaluation of the product.

In such subjects as typewriting, handwriting, cooking, and many others, *both* product and process can be evaluated. In such cases, we prefer to evaluate the products. Evaluation of products tends to be more reliable than evaluation of processes for a number of reasons:

1. More time is available for judging products, while performance in process must be judged "on the wing."
2. Independent judgments of products, made by different evaluators, can be obtained, checked for interscorer reliability, and combined.
3. We can develop a scale of products (representing approximately equal differences in quality) to aid in objective scoring.
4. We can train persons in the use of product scales; we can check their reliability in grading and the extent to which their judgments agree with those of adjudged experts.

In some situations, early errors can irrevocably influence a product; that is, unless the product is evaluated *at different stages*, we can score a product too low because of early errors (for example, an irremediable error in cutting a garment).

Ranking Processes and Products

When a teacher assigns grades to students in such performance skills as swimming or playing tennis, or when he grades their products (whether they are essays or pies), he is presumably ranking the performances or products on some continuum of quality. Too often the characteristics used as a basis for grading differ from teacher to teacher, and also differ as the same teacher observes and evaluates the work of different students. One advantage of using checklists and rating scales, discussed in the next chapter section, is to minimize these differences with respect to selective attention and recall.

Another problem in grading is differences in generosity (from teacher to teacher, and from time to time with the same teacher). If all teachers rank students' work, differences in generosity do not affect students' scores; no teacher can place an unusually large number in the top 10 percent; whereas he could be very generous in assigning A's. Moreover, the ranking process requires the teacher to make a more careful study of interindividual differences than is usually considered necessary in the assignment of marks.

Diederich's suggestion for the sorting of essays into nine groups, as a basis for assigning stanine scores, is a less arduous procedure than placing

all products in rank order. The following plan has been used by assistants or readers who have been trained to grade themes for teachers.

The readers first sort the papers into five piles in order of merit, with 10%, 30%, 20%, 30%, and 10% of the papers in each pile from low to high. Then they take the piles above and below the mean and sort them again in the ratio of *two papers to three*. Thus the first pile of 10% becomes two piles with 4% at the very bottom and 6% slightly better. The next pile of 30% becomes two piles with 12% worse papers and 18% better ones. The same proportions are observed for the two piles above the mean, so that we come out with nine piles with 4%, 6%, 12%, 18%, 20%, 18%, 12%, 6%, and 4% in ascending order of merit . . . a very slight rounding error [has been made] at two points in the scale to make the proportions easy to remember and compute, but they are extremely close to the true stanine proportions.⁵

This procedure could be used with any other products that could be sorted, such as drawings, blueprints, maps, photographs, or small work samples made in industrial arts or home economics.

Using Checklists and Rating Scales

As aids to the observers of processes, and the judges of products, checklists or rating scales should be developed. A checklist merely provides a systematic basis for recording observational data. A rating scale differs from a checklist in that qualitative judgments are made and recorded.

USING CHECKLISTS A checklist is an aid to the observer in recording information regarding sequence of acts or use of approved methods. The person using the checklist might simply check the actions that occur or the methods used. Or he might fill in numbers to indicate the sequence of actions; the completed checklist then constitutes a step-by-step summary of the students' procedures.

A checklist might be used to record those elements in a complex task that had been satisfactorily completed. In scoring the dissection skills, tested in the performance test on page 404, the results of each student's work is checked against a prepared list of the characteristics that the dissection would show if it were properly done. Ideally, this checklist should be developed by the students, or at least discussed with them, before it is used to appraise their work.

⁵ Paul B. Diederich, "Simplified Measurement Techniques for Teachers," *The 15th Yearbook*, National Council on Measurements Used in Education (New York: The Council, 1958), p. 25.

SCORING FOR DISSECTION

- a. Is the skin completely removed from the leg and foot? (1) _____
- b. Are the muscles, tendons, and joints uninjured and intact? (1) _____

Score the remaining items for the gastrocnemius muscle and the tibialis anticus muscle separately. Allow the specified number of points credit for each muscle.

	Gast.	Tib. Ant.
c. Is the muscle completely separated from adjacent muscles?	(2) _____	_____
d. Is the muscle attached at the origin?	(1) _____	_____
e. Is the muscle fully dissected at the origin, its attachment distinct?	(1) _____	_____
f. Is the muscle attached at the insertion?	(1) _____	_____
g. Is the muscle fully dissected at the insertion, its attachment distinct?	(1) _____	_____
h. Is fascia of muscle smooth, not torn?	(2) _____	_____
i. Are the fiber bundles entire, not frayed out?	(1) _____	_____

Score on Dissection

Sum of items a to i inclusive

Students whose dissections are poor can be observed individually as they repeat their work. Errors in procedure can be checked against a teacher-made list.⁶

USING RATING SCALES A rating scale, unlike the checklist, requires a *qualitative* evaluation of aspects of a total performance or product, or of steps or subtasks within a series. The first step in constructing a rating scale is to break down the process or product into components. Decisions may also be made concerning the relative importance of different components.

The rating scale in Figure 12.1 illustrates the rating of different *steps* of a performance in process, that is, sawing to a line with a rip and cross-cut saw. The rating scale in Figure 12.2 illustrates the rating of different *aspects* of a process. Both these types of rating scales are more useful in diagnosis than an over-all rating of the process or product as a whole. For example, a student might receive a perfect score on all aspects of fastening screws except items 3 and 7. If a total score on all items were computed, it would be high; yet a student who hopelessly splits the wood into which the screw is driven needs further teaching and practice.

Rating scales also differ with respect to type of scale used. In Figures 12.1 and 12.2, a simple numerical scale is used. In Figure 12.4 each numerical value is verbally defined in such a way as to encourage instructors

⁶ For a suggested checklist for scoring the dissection performance in process, see Hawkes, Lindquist, and Mann, *op. cit.*, pp. 254-255.

TO SAW TO A LINE WITH A RIP AND CROSS-CUT SAW

Tools and Materials: Sharp rip saw and cross-cut saw, bench, wood vise, and piece of wood.

Directions: Observe pupil as he works, and rate him on the following points:

-
- | | | | | | | | | | | |
|--------------------|---|---|---|---|---|---|---|---|---|----|
| 1. CLAMPING STOCK: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|--------------------|---|---|---|---|---|---|---|---|---|----|
- Stock should be so held that it will not be loosened or cracked, and that its position will facilitate sawing.
- | | | | | | | | | | | |
|------------------|---|---|---|---|---|---|---|---|---|----|
| 2. STARTING CUT: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|------------------|---|---|---|---|---|---|---|---|---|----|
- With thumb at line, saw should be placed against the thumb. Saw should be pulled back slowly a few times to make a groove, then pushed forward.
- | | | | | | | | | | | |
|-----------------|---|---|---|---|---|---|---|---|---|----|
| 3. HOLDING SAW: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-----------------|---|---|---|---|---|---|---|---|---|----|
- Saw should be held firmly. For cross-cut saw, angle should be 45 degrees; for rip saw, 60 degrees.
- | | | | | | | | | | | |
|------------|---|---|---|---|---|---|---|---|---|----|
| 4. STROKE: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|------------|---|---|---|---|---|---|---|---|---|----|
- Stroke should be long and even, not too fast. Proper angle should be kept during sawing. Line should be followed.
- | | | | | | | | | | | |
|----------------|---|---|---|---|---|---|---|---|---|----|
| 5. ENDING CUT: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----------------|---|---|---|---|---|---|---|---|---|----|
- The piece being cut off should be held with the free hand. Saw strokes should be slow and with little pressure so as to prevent breaking off the end.

Fig. 12.1. Form for Rating Different Steps in a Process. (A rough point-scale for judging ability to saw to a line with a rip and cross-cut saw.)

Reprinted by permission of the publisher from M. M. Proffitt, and others, "The Measurement of Understanding in Industrial Arts," *The Measurement of Understanding, 45th Yearbook of the National Society for the Study of Education, Part I* (Chicago: National Society for the Study of Education, 1946), pp. 302-320.

to distribute scores more widely than such terms as "superior," "good," and "fair"; in Figure 12.3 an attempt is made to define the extremes of each scale in terms of observable characteristics.

Figure 12.3 illustrates good procedure in devising a rating form for products. That is, good and poor products have been compared and those characteristics that differentiate them have been identified and included in

(a) NAILS

- | | |
|------------------|----------------------------------------------------------------------------|
| (1) Straightness | 1 2 3 4 5 6 7 8 9 10 |
| | Are nails driven straight, heads square with wood, no evidence of bending? |
| (2) Hammer marks | 1 2 3 4 5 6 7 8 9 10 |
| | Is wood free of hammer marks around nails? |
| (3) Splitting | 1 2 3 4 5 6 7 8 9 10 |
| | Is wood free of splits radiating from nail holes? |
| (4) Depth | 1 2 3 4 5 6 7 8 9 10 |
| | Are depths of nails uniform and of pleasing appearance? |
| (5) Spacing | 1 2 3 4 5 6 7 8 9 10 |
| | Are nails spaced too close or too far apart? |
| (6) Utility | 1 2 3 4 5 6 7 8 9 10 |
| | Will the nails hold? |

(b) SCREWS

- | | |
|------------------------|--------------------------------------------------------------------|
| (1) Slots | 1 2 3 4 5 6 7 8 9 10 |
| | Are slots free of splitting and other evidence of driving strains? |
| (2) Straightness | 1 2 3 4 5 6 7 8 9 10 |
| | Are screws straight, heads parallel with surface? |
| (3) Splitting | 1 2 3 4 5 6 7 8 9 10 |
| | Is wood free of splits in the area of screws? |
| (4) Screw driver marks | 1 2 3 4 5 6 7 8 9 10 |
| | Is wood free of screw driver marks near screws? |
| (5) Countersinking | 1 2 3 4 5 6 7 8 9 10 |
| | Is countersinking neat and of satisfactory depth? |
| (6) Spacing | 1 2 3 4 5 6 7 8 9 10 |
| | Are screws spaced too close or too far apart? |
| (7) Utility | 1 2 3 4 5 6 7 8 9 10 |
| | Will the screws hold? |

Fig. 12.2. Form for Rating Different Aspects of a Process. (Point-scale rating form for "fastening" in woodworking.)

From D. C. Adkins, and others, *Construction and Analysis of Achievement Tests* (Washington, D.C.: Government Printing Office, 1948), p. 231.

the rating form. The same procedure is useful in devising a rating scale for judging performance in process. That is, teachers who are designing a rating scale should compare good and poor performers (violinists, baseball pitchers, or performers in any other skill) as a basis for identifying those component skills in which they differ widely. The reader will recognize that this procedure is similar in approach to the selection of test items on which high-achieving and low-achieving students show the greatest difference in performance. Many other suggestions for the improvement of rating forms and rating procedures are given in Chapter 8.

	1	2	3	SCORE
<i>Appearance of White</i>	1. Dull		Soft luster	1. _____
	2. Spread out and irregular		Thick with rounded outline	2. _____
	3. Greasy		No excess fat	3. _____
<i>Appearance of Yolk</i>	4. Broken		Whole	4. _____
	5. Not coated with white		Coated with white	5. _____
<i>Consistency of White</i>	6. Watery or very solid		Uniformly coagulated	6. _____
<i>Tenderness of White</i>	7. Leathery or crisp and hard		Tender	7. _____
<i>Taste and Flavor</i>	8. Stale, flat, salty, or unpleasant fat flavor		Fresh, well seasoned	8. _____
<i>Total Score</i>				_____

Fig. 12.3. Rating Scale for Eggs (Fried)

Reprinted with the permission of the publisher from Clara Brown Army, *Minnesota Food Score Cards* (Princeton, N.J.: Educational Testing Service, 1946).

Using Product Scales

A product scale is a graded series of products (usually five or more) carefully chosen to represent successive levels of quality along an inferior-superior continuum. In the evaluation of handwriting and composition skills, product scales have been used for many years. In fact, the first product scale in handwriting was developed by Thorndike in 1910.

In the development of a product scale, specimen products are selected (on the basis of ratings by experts) as representing different levels of quality; these products are then used as a basis for grading students' work. In order for the products to constitute an equal-interval scale, the difference in quality between specimens A and B should be approximately as great as between specimens B and C, and so on throughout the scale.⁷

Once the scale of products has been developed and scores assigned, it can be used as the basis for evaluating student products. That is, each student's handwriting sample, essay, or other product is given the score of the specimen it most closely resembles in general quality. Product scales in handwriting have proved fairly satisfactory. The use of product scales in essay writing, however, has involved greater subjectivity and consequently lower reliability.

ILLUSTRATIVE TECHNIQUES IN THE EVALUATION OF COMMUNICATION, MANIPULATIVE, AND ATHLETIC SKILLS

Rating scales, product scales, and other evaluation techniques can be used in evaluating products or processes in a wide variety of skills. A few examples will be given to illustrate the wide variety of possible approaches.

Communication Skills

Figure 12.4 illustrates a rating scale used in evaluating communication skills. This form is used by a college student's peers to rate his effectiveness in speaking. Each student critic not only gives an over-all evaluation of the student's speech but also rates it on the basis of the following four criteria: (1) adaptation to the communication situation (assignment, speaker, and audience); (2) structure; (3) developmental materials, (originality, freshness, accuracy, adequacy, and relevancy); and (4) skill in expression (extemporaneous delivery, use of language, use of voice, and use of body). The listing of more detailed behaviors under each criterion not only helps to clarify the criterion for the student critics but may be used as an additional means of communication to the student being

⁷ In the development of such product scales, it is assumed that the larger the percentage of judges observing a difference between two products, the larger the difference. If 80 percent of the judges indicate that product A is better than product B, while only 70 percent rate B as better than C, the difference between products A and B is considered to be greater than that between products B and C. Differences noted by an approximately equal percentage of judges are considered to be of the same size.

rated. That is, student raters could underline statements that are especially applicable to the speaker. For use at the high school level, the scale should be simplified and probably reduced to a five-point rating scale.

Product scales in essay writing have proved sufficiently reliable for evaluating the average level of student competency in a class or school, especially if teachers have been trained in their use.⁸ However, the problem of appraising *individual* proficiency in essay writing is not so easily solved. Students vary in their effectiveness from time to time and from topic to topic. After extensive research, the College Entrance Examination Board discontinued the grading of essays⁹ and substituted objective and semi-objective tests of related skills.

The staff of the Educational Testing Service has done extensive research concerning ways in which essays can be more objectively graded. In 1957, for the first time in several decades, a published test of essay writing of the product-scale type appeared as part of the STEP series. In order to obtain the specimen student essays for the eight essay topics at each level, five thousand student essays were examined and rated independently by experts in the composition skills.

The essay topics were selected to stimulate students to self-expressive, creative writing, rather than routine narration. The following essay topic is for Level 3 (grades 7-9):

If you knew that you were to go blind twenty-four hours from now and that nothing could prevent it, what would you do with the time between now and this moment, this time tomorrow when you could see no more? Where would you go? What and whom would you try to see? Write an account of what you would do from the time you leave this room until your blindness strikes, explaining, if possible, the reasons for your actions.¹⁰

In the week-by-week evaluation of themes, a more individualized appraisal than is possible with product scales is desirable. The way in which the teacher reads and evaluates students' compositions can have lasting effects on their attitude toward writing. The teacher should not limit his notations to proofreading symbols. Comments, directed to the individual writer and based on the teacher's understanding of his needs and capacities, can serve as a stimulus and guide to improvement.

⁸ Harry A. Greene, "English—Language, Grammar, and Composition," *Encyclopedia of Educational Research* (New York: The Macmillan Company, 1950), p. 394.

⁹ Copies of each student's essays, however, are still supplied to the colleges to which he makes application.

¹⁰ *A Prospectus, Cooperative Sequential Tests of Educational Progress* (Princeton N. J.: Educational Testing Service, 1957).

Directions: Indicate your rating on the five aspects of the speech by drawing a circle around the number which represents your rating in each case.

Student _____ Subject _____

I. Adaptation to the Communication Situation

SUPERIOR	EXCELLENT	GOOD	AVERAGE	FAIR	POOR	VERY POOR
14 13	12 11	10 9	8 7	6 5	4 3	2 1

- A. *Suited to the assignment*: follows assignment—stays within set time limits
- B. *Suited to the speaker*: ethical justification—well prepared—desire to communicate
- C. *Suited to the audience*: clear articulation—correct pronunciation—visual aids where necessary—neatly dressed—poised in body and facial expression—eye contact—conversational tone—variation in pitch, rate, and loudness—subject appropriate

II. Structure of the Speech

SUPERIOR	EXCELLENT	GOOD	AVERAGE	FAIR	POOR	VERY POOR
14 13	12 11	10 9	8 7	6 5	4 3	2 1

- A. *Introduction*: captures attention—focuses attention on subject—leads to stated or implied purpose—establishes mood of speech—proper length for balance
- B. *Body*: subject-analysis pattern clear—all divisions subordinate to purpose—each division a separate unit—division by one principle only
- C. *Conclusion*: restates purpose and summarizes—uses additional element (pithy epigram or reference to previous illustration, etc.)—proper length for balance
- D. *Transition elements*: verbal bridges between divisions clear
- E. *Sentences*: correct structure—varied structure—effective parallelism

III. Developmental Materials

SUPERIOR	EXCELLENT	GOOD	AVERAGE	FAIR	POOR	VERY POOR
14 13	12 11	10 9	8 7	6 5	4 3	2 1

- A. *Originality of material*: selection of material from many sources—initiative in gathering data—no hint that the material is paraphrased from book or magazine
- B. *Freshness of material*: uses personal experiences—avoids outdated data—adapts old facts to new contexts
- C. *Accuracy of material*: uses honest details—qualifies opinions—uses specific support—avoids questionable authority
- D. *Adequacy of material*: sufficient details—sufficient illustrative devices—use of statistics or apt quotations when available
- E. *Relevancy of material*: details pertinent—details realistic—connection between examples and generalization demonstrated

IV. Skill in Expression

SUPERIOR		EXCELLENT		GOOD		AVERAGE		FAIR		POOR		VERY POOR	
14	13	12	11	10	9	8	7	6	5	4	3	2	1

- A. *Extemporaneous delivery*: speaks without notes—minimum of vocalized pauses—effective use of unvocalized pause—adaptation to audience reactions—rapport with audience
- B. *Use of language*: avoidance of clichés—sense of sentence rhythm—exactness in word choice—recognition of connotative value of words
- C. *Use of voice*: voice modulated to verbal symbols—pleasant tonal quality
- D. *Use of body*: projects alert body tone—purposive movement—co-ordinated movement—natural and spontaneous gestures

V. Over-all Evaluation

SUPERIOR		EXCELLENT		GOOD		AVERAGE		FAIR		POOR		VERY POOR	
14	13	12	11	10	9	8	7	6	5	4	3	2	1

Name of student making the rating _____

Fig. 12.4. Scale for Evaluating Speaking

Reprinted by permission of The American Council on Education from Paul L. Dressel and Lewis B. Mayhew, *General Education: Explorations in Evaluation* (Washington, D.C.: American Council on Education, 1954), p. 81.

The following summary of "Principles of Theme Analysis," although developed for use at the college level, has many implications for the high-school teacher of composition.

PRINCIPLES OF THEME ANALYSIS¹¹

1. *External motivation*. Read a student's theme in the light of its external motivation—the audience and assignment to which it is addressed—and its success in meeting the requirements of that audience and assignment.
2. *Internal motivation*. Strive to understand the student's internal motivation in writing the theme—what he is trying to do. In order to appreciate the student's purpose, one's attitude in reading must be one of constructive helpfulness rather than negative criticism.
3. *Unrealized potentialities*. Be alert to the unrealized potentialities of the theme, the opportunities wasted or used without imagination. Calling these to the attention of the student will give a valuable stimulus to future writing.

¹¹ Adapted from Paul L. Dressel and Lewis B. Mayhew, "Objectives in Communication," in *General Education: Explorations in Evaluation* (Washington, D.C.: American Council on Education, 1954), pp. 86–89.

4. *Interdependence of parts.* Recognize the interdependence of the parts of a theme. To view a part accurately is to see it in relationship to the whole.
5. *Concluding evaluative judgment.* Relate your concluding judgment specifically to the subject matter of the theme and see that it is consistent with the running commentary. When evaluating a "good" theme, avoid the temptation of limiting your observations to minor flaws and concluding with a congratulatory message. If the good student is to achieve his best, a full and conscientious reading must call attention to both his successes and his failures.

Manipulative Skills

Micheels and Karnes¹² contend that, in determining final course marks in industrial arts, teachers give far more weight to the quality of finished products than to any other factor, frequently giving little consideration to the design and planning of the project and to the procedures followed. The fallacy of this procedure is evident when one considers that a student might eventually complete a project of high quality and yet, in the process of its construction, have done one or more of the following:

1. Consumed an unjustifiable amount of time in the completion of the project.
2. Asked for and obtained more assistance from the instructor and from his fellow-students than any other member of the group.
3. Wasted an undue amount of materials.
4. Performed inaccurate and faulty work which was concealed when the project was assembled.
5. Abused tools and equipment; failed to use them properly.
6. Persistently violated safety rules.
7. Failed to follow the general procedure as initially planned.
8. Failed to accept the challenge to design a project of his own or even select and adapt a design but waited for the instructor to assign him a design to execute.
9. Showed no evidence of having developed an appreciation of good design and skilled workmanship.
10. Failed to learn the related information about tools, materials, and processes which was assigned as a part of his project.¹³

Figure 12.5 presents a comprehensive teacher rating scale for the designing, planning, and executing stages of woodwork projects. The application of such rating scales can be the basis for individualized diagnosis and efforts for improvement. Such ratings, of course, have value only as they

¹² William J. Micheels and M. Roy Karnes, *Measuring Educational Achievement* (New York: McGraw-Hill Book Company, Inc., 1950), p. 398.

¹³ *Ibid.*, p. 399.

are used conscientiously by the instructor, in conjunction with (1) study of the student's drawings, specifications, and plan of procedure, and (2) inspection of the finished project in every detail, using such measuring instruments as are necessary to determine the accuracy and quality of the student's work.

Manipulative skills are also involved in home economics. Product scales (with samples of different quality levels of hand sewing, French seams, bound buttonholes, and the like) can be used to advantage in teacher rating of products, in student self-rating, and in obtaining ratings by peers. In the development of such product scales, students' ratings of products can be utilized. Those samples on which student ratings show high agreement can be selected to represent approximately equal differences in quality.

After products that represent several degrees of quality have been selected and scores assigned to them, each student's product can then be given the score of the specimen it most closely resembles. Independent ratings can be obtained by two or more judges (such as fellow-students) and their values averaged. The teacher should personally appraise any products for which the student's self-evaluation differs from the average rating assigned by classmates.

The homemaking teacher also finds that observation of performance in process is necessary as an aid in diagnosis and as a basis for reteaching. On the basis of a comparison of superior and inferior products in cooking and sewing, she can develop checklists, teacher-rating scales, and self-rating scales that will be of great value in improving students' procedures.¹⁴

Athletic Skills

In physical education both rating scales, and more objectively scored performance tests, are used in evaluating the quality of student performance in the component skills involved in gymnastics, track, such individual sports as tennis and badminton, and team games. Whether rating scales or more objectively scored performance tests are used, the first step is to analyze the component skills involved in a sport. The following analysis for basketball is illustrative:

1. Shooting
 - a. Foul
 - b. Shooting from the floor
 - (1) One hand
 - (2) Two hands
- } Types of one- and two-hand shots

¹⁴ Clara Brown Army, *Evaluation in Home Economics* (New York: Appleton-Century-Crofts, 1953).

Name: _____ Course: _____
 Project: _____ Score: _____
 Instructor: _____ Date: _____
 Numbers of items which do not apply: _____

Directions: Each of the items in this scale is to be rated, if it applies, on the basis of 4 points for outstanding quality, degree, compliance, or performance; 3 points for better than average; 2 points for average; 1 point for inferior; and 0 for unsatisfactory or failure. Encircle the appropriate number to indicate your rating. Draw a horizontal line through the row of numbers opposite each item which does not apply. Enter the total points earned under each major phase. Enter the composite total in the space at the top of this sheet. Also indicate the number of items which do not apply.

I. Designing Phase (Total Points _____)

1. To what extent is the project designed or selected of value to him or to his associates? 0 1 2 3 4
2. To what extent did he evidence sensitivity to the elements of good design?
 - a. Size, proportion, balance, relative weight of parts? 0 1 2 3 4
 - b. Texture, color surface, and line enrichment? 0 1 2 3 4
3. Is the material selected appropriate? . . . 0 1 2 3 4
8. To what extent were his sketches and drawings orderly and generally indicative of good workmanship? 0 1 2 3 4

II. Planning Stage (Total Points _____)

1. Did he obtain the basic information about tools, materials, and processes essential to intelligent planning? 0 1 2 3 4
2. To what extent did he prepare his own plan of procedure? . . . 0 1 2 3 4
6. To what extent did he take into consideration the time, materials, equipment, and tools available? 0 1 2 3 4

III. Execution Stage (Total Points _____)

1. To what extent did he follow the detailed steps of his plan? 0 1 2 3 4
2. To what extent did he avoid having to do work over because of failure to follow his plan? 0 1 2 3 4
3. To what extent did he refrain from spoiling materials by working accurately and carefully? 0 1 2 3 4
4. To what extent did he follow approved procedures in performing specific operations? . . . 0 1 2 3 4
13. To what extent was he able to do his own work without assistance from instructor or other students? 0 1 2 3 4

IV. Completed Project (Total Points _____)

1. To what extent is finished product an embodiment of original plan?	0	1	2	3	4
2. Does the general appearance of the project reflect neat, orderly work?	0	1	2	3	4
3. Are the dimensions of the actual project the same as those on the drawing, within reasonable tolerances?	0	1	2	3	4
4. How do angular measurements check with those specified?	0	1	2	3	4
5. Of what quality is the finish?	0	1	2	3	4
6. To what extent were materials used to best advantage?	0	1	2	3	4
7. Do all joints fit properly?	0	1	2	3	4
8. Are all margins uniform? are curved and irregular lines properly executed, etc.?	0	1	2	3	4

Fig. 12.5. A Teacher Rating Scale for the Designing, Planning, and Executing Stages of Woodwork Projects

Reprinted by permission of the publisher from William J. Mischeels and M. Roy Karnes, *Measuring Educational Achievement* (New York: McGraw-Hill Book Company, Inc., 1950), pp. 408-410.

2. Ball handling
 - a. Passing (subdivided into different kinds of passes)
 - b. Receiving
 - c. Dribbling (and combinations)
3. Total body skills
 - a. Jumping
 - b. Speed
 - c. Pivot
 - d. Endurance¹⁵

A number of available tests of sports skills, suitable for use in secondary schools, are listed in Clarke¹⁶ and in Adams and Torgerson.¹⁷ Since the

¹⁵ Leonard A. Larson and Rachel D. Yocom, *Measurement and Evaluation in Physical, Health and Recreation Education* (St. Louis: The C. V. Mosby Company, 1951), pp. 208-209.

¹⁶ H. Harrison Clarke, *Application of Measurement to Health and Physical Education*, 3d ed. (Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1959).

¹⁷ Georgia Sachs Adams and T. L. Torgerson, *Measurement and Evaluation for the Secondary School Teacher* (New York: Holt, Rinehart and Winston, Inc., 1956), pp. 466-483.

better skills tests require considerable time to administer, they tend to be used only with activities in which instruction has been given for several weeks. Rating scales are more frequently used for activities receiving less emphasis and for all those skills (such as dancing or swimming) in which grace, balance, and form are emphasized.

VALIDITY AND RELIABILITY OF EVALUATIONS OF STUDENT PERFORMANCE IN THE SKILLS

Validity

Since we have been concerned in this chapter with the *direct* measurement of actual student behavior and actual products made, it may seem that no problems of validity are involved and that we may concentrate on reliability of measurement. This would be the situation if our sampling of each student's behavior were highly representative of his work as a whole; and our scoring of that sample was unaffected by extraneous factors. When products can be scored objectively, such as the speed and number of errors in typewriting, we can take validity for granted. However, few of the evaluation situations described in this chapter are entirely free from the bias of raters or from extraneous factors that obscure our evaluation of the criterion behavior in which we are really interested.

We can increase the validity of our ratings of performance in process by using recording techniques. If we can replay a record of a student's pronunciation in foreign language or view a film (in slow motion) of his form in a swimming stroke, the validity of our observation and scoring is increased. We are less likely to have our ratings affected by our general impression of the student (his helpfulness and his evidences of interest in the work). Also we can check our judgments against those of another qualified judge.

If we can listen to recordings, view films of student performance, or rank student products without knowing whose work we are grading, validity is increased because the "halo effect" of our general impression of the student is eliminated. If we want to know whether students have improved, *on the average*, in their handwriting, pronunciation, or essay writing, we could intermix student products taken during the first and last months of the year and then rank or rate a sampling of these without any knowledge of whether each product was obtained in the fall or spring of the year. This would eliminate the effect of wishful thinking and increase the validity of our inferences concerning how much the *group* had progressed.

We also increase the validity of scores on skills when we make sure that

extraneous factors have minimal effect on student performance. For example, we should have students use the same kinds of equipment (in equally good repair) and the same kinds of materials (lumber, dress material, and the like) if we are to make valid comparisons with respect to their relative skill.

When subjective judgments must be made by observers, validity can be increased by (1) selecting competent observers who know the crucial elements of a good performance or product and by (2) training them to observe those aspects of a performance or product that most clearly differentiate students who are rated high or low on much larger samplings of performances or products.

Reliability

Evaluation in the skills presents difficult problems in achieving an adequate level of reliability for interindividual comparisons. A review of Chapter 4 reveals that the major factors affecting reliability of scores are (1) length of test or size of sample, (2) objectivity in scoring, (3) consistency in test administration, and (4) appropriateness of the level of difficulty for students. Of these, the last two factors are ordinarily easier to cope with than the first two. We have already considered in this chapter the desirability of the third factor, that is, having clear, standard directions.

As far as the fourth factor, difficulty of the test, is concerned, students can be most fairly compared if they are all given the same test. In many skills, such as typing, running, and the like, the proficiency of all students can be adequately measured in a test situation suitable for all. In other situations, we would have a more efficient and reliable test of skill if students were grouped and the difficulty of the test adjusted for lower-ability and higher-ability students. However, unless we had some way of obtaining comparable converted scores (for example, by giving all tests to a sampling of the student population), we would have sacrificed comparability of scores.

It makes good sense, for example, to test our most able instrumental music students on the most difficult selections. Such a test does not waste time in having them play selections that are too easy for them and it provides a more efficient adequate basis for *differentiating among* the more able students. The teacher may choose these gains and give up the advantages of the uniform test content. Or the teacher may utilize the concept of a uniform "anchor test" for all students plus (a) more difficult musical selections for the most proficient and (b) selections of less-than-average difficulty for the least proficient. The more accurately the teacher can peak the test to the student's level of competency, the more adequately the stu-

dent can demonstrate his skill *in a limited amount of testing time*; and the more reliable his test results are likely to be on a test-retest basis.¹⁸

The first factor, size of sample, is a genuine problem in those tests of skill where a scorable unit requires considerable student time; that is, so much time is required for a student to write one essay, bake one cake, or make one blueprint that one cannot include nearly as many items as on a paper-and-pencil test. Sometimes one can select crucial elements of a task (as in the writing tasks on page 389 of Chapter 11). Or, as was suggested on page 403, one could increase the number of tasks administrable in a given testing time by having them partially completed. The answer to the problem of limited sampling varies from area to area. Certainly, one possibility is to cumulate products, or stanine scores on work samples, over a period of time. Then letter grades could be assigned on the basis of such cumulated data.

The problem of improving objectivity of scoring is involved in the evaluation of most skills. There are intrinsic differences from skill to skill in the extent to which one must depend on subjective judgment. For example, in the area of physical education, speed of running or height in the high jump or rope climb can be scored with as high reliability as we obtain in objective tests. Students' relative skill in body control, as shown in the head stand, or their steadiness in the hand stand can be judged with only moderate reliability, because we must depend to some degree on the subjective judgment of observers. Evaluation of a player's form in tennis or swimming is affected even more by subjectivity in judgment.

In general, reliability is increased by narrowing the basis for ranking or rating and clearly defining the characteristics of a good performance or product. For example, when we independently rate specific, clearly defined aspects of sawing, fastening, or some other skill or product, reliability of scores increases. When we rank the *general* merit of a performance or product, reliability tends to lower. This drop in reliability is due to lack of clarity with respect to (1) the characteristics to be considered and (2) the relative weight to be assigned to them. When one ranks essays, for example, what characteristics are to be considered? How much weight is to be given to neatness, correctness of spelling, organization, originality, and other factors? Checklists or rating scales help to increase interscorer reliability by providing partial answers to such questions. Diederich¹⁹ developed rating scales for judging essays with respect to (1) organization, (2) style, (3) logical reasoning, and (4) content. Product scales can be used for

¹⁸ For further discussion of the dilemma of reconciling the needs for (1) a wide spread of test difficulty and (2) a limited amount of testing time, the student is referred to the discussion of the *SRA Achievement Series* in Chapter 13.

¹⁹ Paul B. Diederich, "Measurement of Skill in Writing," *School Review*, vol. 54 (December 1946), pp. 584-592.

different aspects of a product; for example, handwriting can be rated for beauty and legibility.

Reliability can be increased by averaging judgments made by two, three, or more persons who make their judgments independently (without knowledge of ratings assigned by others). With older students, who understand the importance of learning to evaluate products, the teacher can use a systematic plan for having each student's products (with no name attached) evaluated by several of his classmates.²⁰

Another source of error variance is instability in the performance itself, or student inconsistency on different testing occasions. Actually, it is only when we get reasonable interscorer agreement and reasonably standard conditions of work that we have any chance to study this source of error variance, that is, the variation in student performance from one test sample to another. Traxler and Anderson²¹ had students write a pair of two-hour essays a few days apart on highly similar topics, "The Discovery of Gold in California" and "The Pony Express." They distributed a set of instructions for writing the essay, an outline specifying the four main divisions of the paper, and a set of unorganized notes as background information. Each student was required to base his paper entirely on the notes provided. Competent, trained readers showed high interscorer agreement on grades. Yet, despite the care the researchers had taken to make the tasks comparable and to minimize invalid sources of variance, students' scores on the two sets of essays correlated only .60. It seems that students do vary more in their essay writing performance than in their performance on tests of arithmetic or spelling.

Reliability, as well as validity, can be increased by selecting raters who are especially competent in judging certain *aspects* of a performance or product. French made a factor analysis of readers' ratings of essays. He identified a large group of readers who emphasized "Mechanics and Wording." Interscorer reliability for their ratings of student essays was .70. However, the essay scores assigned by these raters correlated so highly with students' scores on objective tests of writing skills that the Board decided to use the objective tests only. They made a further study of students' factor scores on "Ideas," an aspect of writing ability that can be measured only through actual essay writing. Such scores were almost completely unrelated to the "Mechanics and Wording" scores.

²⁰ Simpson developed an ingenious plan for the rating of products by classmates so that each student would have to rate only five products and yet each student's product would be compared with a random sampling of 20 others. The plan is presented in Ray H. Simpson, "Patterns for Rating Learning Products," *Educational and Psychological Measurements*, vol. 13 (Winter 1953), pp. 614-617.

²¹ A. E. Traxler and H. A. Anderson, "The Reliability of An Essay Test in English," *School Review*, vol. 43 (September 1935), pp. 534-539.

Although the "Ideas" score seemed potentially valuable, the reliability coefficient was extremely low, only .31. However, when they selected readers who assigned considerable importance to "Ideas," the interscorer reliability coefficient increased to .46. Although this reliability is still too low for individual scores to be dependable, the increase from .31 and .46 seems to be attributable to the fact that some persons are more competent than others in judging this important and elusive aspect of essay writing.

Since those aspects of essay writing that could be scored with fairly high reliability can be measured even more consistently by objective and semiobjective tests, French recommends that further research be conducted to improve the effectiveness with which we can measure those aspects of essay writing for which objective tests provide no substitute. For example, if the topics were selected and the directions worded so that students would concentrate on certain aspects of essay writing (for example, organization and ideas), reliability of these scores might be further increased. According to French,

If we psychometricians can encourage testing and further clarification of those aspects of writing that objective tests cannot measure, encourage the use of readers who favor grading those particular qualities that are desirable to grade, and see to it that the students are aware of what they are being graded on, we can enlighten rather than merely disparage the polemic art of essay testing.²²

This statement illustrates how objective testing and the subjective judgment of competent judges can supplement each other effectively in evaluating student progress toward major educational goals.

SUMMARY STATEMENT

Evaluation of student performance in the skills outcomes of many subject areas has been neglected because of the inherent difficulties involved. Such tests as have been developed of "performance in process" tend to be of the work sample or simulated-situation types. The tasks selected for performance tests should be tasks crucial to the attainment of major course objectives and ones which are fairly difficult, not too time consuming, and capable of being administered under fairly standard conditions and evaluated with considerable objectivity.

Evaluation of the products of performance tends to be more reliable than evaluation of performance in process since more time is available for the

²² John W. French, "Schools of Thought in Judging Excellence of English Themes," *Proceedings of the 1961 Invitational Conference on Testing Problems* (Princeton, N. J.: Educational Testing Service, 1962), p. 28.

judging process; independent judgments can be made and compared; product scales can be developed; and teachers can be trained in their use. If products are ranked, and/or stanine scores are assigned to products, differences in rater generosity do not affect students' scores.

Checklists can be used to summarize information on the methods used by students, or on the sequences of steps employed. Rating scales require a qualitative evaluation of the performance or product. Ratings obtained on different steps in a process, or on different aspects of a process or product, are more useful in diagnosis than over-all quality ratings.

A number of suggestions were made for increasing the validity and reliability of judgments concerning student performance in the skills. The problems involved, and hence the best techniques of evaluation, vary considerably from one subject area to another. In all areas, however, we are concerned that the sampling of skills be as large and representative as feasible; that aids be provided to increase the fairness and consistency of grading; that student performance be observed or student products obtained under comparable conditions; and that the tasks assigned be of appropriate difficulty.

SELECTED REFERENCES

- ADKINS, DOROTHY C., "Principles Underlying Observational Techniques of Evaluation," *Educational and Psychological Measurement*, vol. 11 (Spring 1951), pp. 29-51.
- AMERICAN ASSOCIATION FOR HEALTH, PHYSICAL EDUCATION AND RECREATION, *Youth Fitness Test Manual*. Washington, D.C.: The Association, 1958.
- ANDERSON, C. C., "The New Step Essay Test as a Measure of Composition Ability," *Educational and Psychological Measurement*, vol. 20 (Spring 1960), pp. 95-102.
- ARNY, CLARA B., *Evaluation in Home Economics*. New York: Appleton-Century-Crofts, 1953, Chapter 7.
- BEAN, KENNETH L., *Construction of Educational and Personnel Tests*. New York: McGraw-Hill Book Company, Inc., 1953, Chapter 6.
- BAKAN, EDWARD E., "How Do V-Ag Graduates Perform?", *Agricultural Education Magazine*, vol. 29 (May 1957), pp. 259, 261-262.
- FRENCH, ESTHER L., AND EVELYN STALTER, "Study of Skill Tests in Badminton for College Women," *Research Quarterly of the American Association for Health, Physical Education, and Recreation*, vol. 20 (October 1949), pp. 257-272.
- GREENE, EDWARD B., *Measurements of Human Behavior*, rev. ed. New York: The Odyssey Press, Inc., 1952, Chapter 9.
- HENDRICKS, B. CLIFFORD, "Laboratory Performance Tests in Chemistry," *Journal of Chemical Education*, vol. 27 (June 1950), pp. 309-311.
- KORAN, SIDNEY W., "Performance Testing in Public Personnel Selection," *Educational and Psychological Measurement*, vol. 1 (July, October 1941), pp. 233-252, 365-386.
- MCPHERSON, MARION W., "A Method of Objectively Measuring Shop Performance," *Journal of Applied Psychology*, vol. 29 (February 1945), pp. 22-26.
- MICHEELS, WILLIAM J., AND M. ROY KARNES, *Measuring Educational Achievement*. New York: McGraw-Hill Book Company, Inc., 1950.

- MILLER, FRANCES A., "A Badminton Wall Volley Test," *Research Quarterly of the American Association for Health, Physical Education, and Recreation*, vol. 22 (May 1951), pp. 208-213.
- MILLER, RICHARD T., "A New System of Tennis Stroke Analysis," *Athletic Journal*, vol. 32 (March 1952), pp. 45-46, 75-77.
- PEAK, HELEN, "Problems of Objective Observation," in Leon Festinger and Daniel Katz, eds., *Research Methods in the Behavioral Sciences*. New York: Holt, Rinehart and Winston, Inc., 1953, pp. 243-299.
- ROTHROCK, THURSTON M., "Checking the Student's Knowledge with the Camera," *Industrial Arts and Vocational Education*, vol. 38 (January 1949), pp. 19-22.
- RYANS, DAVID G., AND NORMAN FREDERICKSEN, "Performance Tests of Educational Achievement," in E. F. Lindquist, ed., *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, Chapter 12.
- SIMPSON, RAY H., "Patterns of Rating Learning Products," *Educational and Psychological Measurement*, vol. 13 (Winter 1953), pp. 614-617.
- SIRO, EINAR E., "Performance Tests and Objective Observation," *Industrial Arts and Vocational Education*, vol. 32 (April 1943), pp. 162-165.
- THORNDIKE, ROBERT L., *Personnel Selection; Test and Measurement Techniques*. New York: John Wiley and Sons, Inc., 1949, Chapters 1-3.
- WALL, CLIFFORD NATHAN, H. KRUGLAK, AND L. E. H. TRAINOR, "Laboratory Performance Tests at the University of Minnesota," *American Journal of Physics*, vol. 19 (December 1951), pp. 546-555.
- WATKINS, JOHN C., "Objective Measurement of Instrumental Performance," *Teachers College Record*, vol. 44 (February 1943), pp. 376-377.
- WRIGHTSTONE, J. WAYNE, *Measuring the Effectiveness of Instruction in Vocational Education*. Albany, N.Y.: University of the State of New York, February 23, 1951.
- , "Observational Techniques," *Encyclopedia of Educational Research*, 3d ed., C. W. Harris, ed. New York: The Macmillan Company, 1960, pp. 927-933.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. Discuss the values and limitations of tests of "performance in process" in homemaking, industrial arts, or some other subject in which the development of skills is emphasized.
2. Discuss the values and limitations of product scales in the same subject field.
3. Outline your plans for developing a performance test of the work sample or simulated-situation type for use in evaluating skills in your major subject field. Follow the guidelines that are presented in this chapter for selecting tasks for performance tests.
4. Develop a checklist for recording the sequence of acts or the use of approved methods in the performance of some manipulative skill.
5. Discuss the values and limitations of rating scales in evaluating student skills in a specific sports activity, for example, tennis or diving.
6. Prepare a guide to be used in appraising students' laboratory skills, as shown in a specific laboratory exercise.

7. Obtain several specimens of handwriting, and evaluate them on the basis of a handwriting scale of the survey type.
8. Assume that you are head of an English department; prepare a bulletin of suggestions for teachers of composition on the evaluation of students' themes.
9. Modify the speech rating scale (Fig. 12.4) so as to make it more suitable for use on the high school level.

The Place of Standardized Achievement Tests in the Improvement of Instruction

Before we consider the current place of standardized achievement testing in education, we will trace briefly the history of such testing in the American schools. Such a review will help us to understand some of the contributions made by this type of testing, as well as some of the reasons why educators are divided as to whether standardized tests constitute aids to achieving educational goals or whether they are a barrier to significant progress.

Actually, most educators have come to realize that standardized tests can be either helpful or harmful, depending on the wisdom with which they are selected and used. In the barrage of criticism against testing, the most valid criticisms have had to do with the inadequacies of specific tests or with the misinterpretation and misuse of test data by fallible human beings. The basic approach involved in measuring achievement through a wide variety of standardized tests, among which test users may make a choice, has held up well under criticism.

HISTORY OF ACHIEVEMENT TESTING

When one considers the widespread use of standardized tests today, it is difficult to realize the youth of objective testing. In fact, it was only about a century ago that school enrollments became sufficiently large that uniform written examinations were first adopted in the schools of Boston as a substitute for the characteristic oral examinations used to pass on the qualifications of students.¹

¹ Robert C. Hall, "Types of Tests Available," *School Life*, vol. 42 (September 1959), pp. 10-13.

The Beginnings of Standardized Achievement Testing

Just as group intelligence tests were developed in order to meet a practical problem, so the need for group achievement tests arose from a practical school situation. As a school administrator, Rice was faced in 1894 with considerable pressure to bring into the curriculum such new, practical subjects as manual training and home economics and also with considerable opposition on the part of educators who thought that there was hardly sufficient time to teach the subjects already in the curriculum. As an initial step toward studying scientifically the effectiveness of instruction under different time allotments, Rice decided to administer uniform tests in spelling in a number of schools.

The next achievement tests developed were the *Stone Reasoning Test in Arithmetic*, published in 1908, and the *Thorndike Handwriting Scale*, published in 1909. Because of the number and significance of his contributions in the following decade, E. L. Thorndike is generally considered to be the father of the educational-measurement movement.

Beginning in 1910, a number of studies were made on the unreliability of teachers' grading of students. The findings stimulated the development of more objective procedures for testing the achievement of students and for assigning marks or grades. In one of the most striking of the early studies,² copies of the same geometry paper were marked by 116 teachers of high school mathematics; the grades assigned varied from 28 to 92. Evidence from studies of English composition and other subjects revealed similar inconsistencies. College teachers were found to be shockingly inconsistent when they *regraded papers of their own students* without knowledge of the marks they had formerly assigned.³

Such findings gave tremendous impetus to the development and use of achievement tests that utilized the objective type of questions, developed for use in group intelligence tests during World War I. The need had been established, and the techniques had been introduced. As the decade of the 1920s opened, the stage was set for large-scale development of group tests of achievement.

An important development of the early 1920s was the organization of tests into batteries. In 1922 the first edition of the *Stanford Achievement Test* appeared. In revised forms, this has continued to be one of the leading achievement-test batteries. By administration of a test battery (which includes subtests on the skills of reading, arithmetic, language, and other subjects), measures could be obtained of children's *comparative* achieve-

² Daniel Starch and Edward C. Elliott, "Reliability of Grading Work in Mathematics," *School Review*, vol. 21 (April 1913), pp. 254-259.

³ Daniel Starch, "Reliability and Distribution of Grades," *Science*, vol. 38 (October 1913), pp. 630-636.

ment in these different areas; and the achievement of class and school groups could be interpreted in comparison with age or grade norms (the average achievement for children of the same age or grade level).

Use of the New Tests in School Surveys

Administrators and supervisors soon accepted the new testing techniques as valuable tools by which to compare the achievement of classes and to rate the efficiency of teachers. City, county, and state surveys of school systems flourished during the 1920s. In these surveys, group tests of intelligence and achievement were used extensively in an effort to determine the efficiency of instruction and to study other administrative, supervisory, and curricular problems. The average scores obtained by classes, grades, and schools were compared with national norms for the tests. Objectivity and efficiency in education were emphasized.

This period was also characterized by increased interest in the range of individual differences revealed by testing programs, and in educational research, especially of the type related to determining the efficiency of school services. On the debit side, however, there must be recognized a tendency toward overconfidence in tests and frequent misuse of test results as a basis for judging the quality of teaching. Moreover, the growing market stimulated overproduction of tests, many of which were inferior in quality and in standardization.

Reactions to Criticisms of Standardized Tests

The 1930s and 1940s were characterized by a general acceptance of the usefulness of tests in the schools and the development of a more critical attitude regarding the values of specific tests. Leaders of the progressive movement in education emphasized the fact that tests were concerned almost exclusively with elementary skills and with facts to be memorized. Critics of standardized tests reminded teachers that these tests did not measure progress toward the ultimate goals of education—understandings, attitudes, and appreciations. It was pointed out that a composite of the results of currently used tests for a given individual did not provide a true or complete picture of the individual as a whole.

Such criticisms led to attempts to obtain more comprehensive appraisals through the use of anecdotal records, interviews, and case histories, as well as the development of tests concerned with higher-order cognitive learnings, which go beyond information and skills. Use of these measures tended to provide a better evaluation of the child in terms of his many-sided patterns of development.

Summary of Historical Trends in Achievement Testing

Although space does not permit us to include many of the interesting milestones in the development of achievement testing, the following generalizations will provide background for the study of current uses of achievement tests.

1. The testing movement helped to arouse the profession to the extent and significance of individual differences in student achievement and readiness for new learnings.
2. The inadequacy of early tests and the misuse of test results led to certain undesirable outcomes: (a) standardized tests were frequently used in schools without consideration of their appropriateness in the local educational program; (b) results of survey testing were frequently misused as a basis for judging teaching efficiency; (c) as a result, the teaching in many schools became largely the coaching of students for test passing; and (d) since tests failed to evaluate student growth on a sufficiently broad basis, teachers' emphasis on tested educational outcomes led to an undesirable narrowing of the educational program. What was most easily measured became most important. It is not surprising that many teachers resented the use of standardized achievement tests and that antagonism toward testing developed on the part of many educators who were striving to broaden and vitalize the educational program.
3. Curriculum change and a growing emphasis on child study led to needed modifications in measurement and evaluation. Many standardized tests, which had measured information only, were broadened so as to measure student understandings and application of principles.
4. There has been increased awareness among administrators of the fact that standardized tests measure student growth toward only a limited number of educational goals. Hence greater caution is being exercised in making inferences from test data concerning the *general* teaching effectiveness of faculty members. Administrators have also learned to take into account in their interpretations of test data for classes and schools the many factors that affect school achievement, such as the scholastic aptitude and cultural background of students.
5. As it became apparent that published tests could provide only a partial answer to the problems of appraising student growth, teachers were given preservice and inservice education in the development of their own tests and in the use of other techniques of assessing student growth toward educational goals. The first book designed to help teachers in improving their own examinations was written in 1924.⁴ In the 1930s the work of the Eight-Year Study⁵ in secondary schools provided a tremendous stimulus to the development of test exercises that went far beyond the testing of knowledge and included most of the major objectives listed in the taxonomy,

⁴ G. M. Ruch, *The Improvement of the Written Examination* (Chicago: Scott, Foresman and Company, 1924).

⁵ Eugene R. Smith, Ralph W. Tyler, and others, *Appraising and Recording Student Progress* (New York: Harper & Row, Publishers, Inc., 1942).

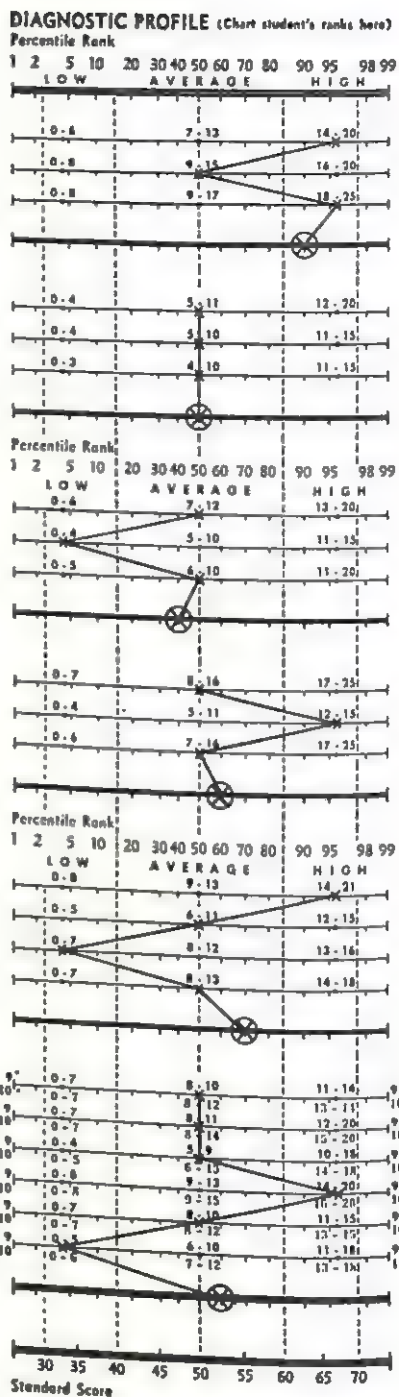
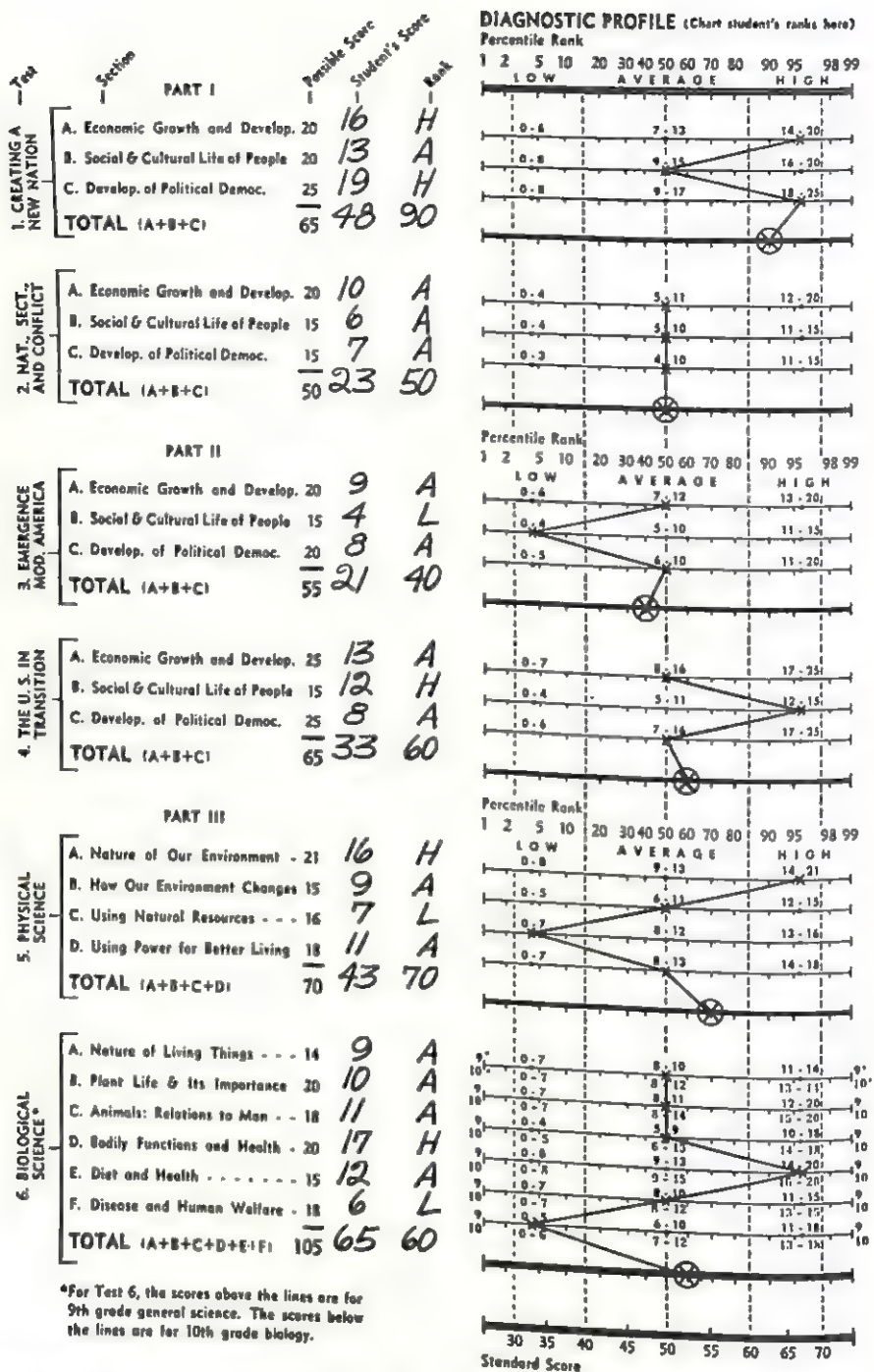


Fig. 13.1 Sample Profile for the California Tests in Social and Related Sciences, Advanced Battery, Administered to an Eleventh-grade Student in May.

Reproduced with the permission of The California Test Bureau (Monterey, Calif.: California Test Bureau, 1954).

Aid in Student Decisions

Problems 6 and 7 in Chapter 1 had to do with the advisement of students regarding the choice of college-preparatory subjects and of vocations. Achievement tests of the generalized-outcome type would be an aid to counselors in these situations. Such tests correlate highly with scholastic aptitude and are much more easily discussed with students and parents than are scholastic aptitude tests.

LEADING ACHIEVEMENT TESTS

With the exception of end-of-course examinations in high school subject fields, schools generally prefer to administer achievement test batteries rather than separate subject tests in reading, arithmetic, language and other areas. The chief reason for their preference is that test batteries provide comparable scores for students in all subtests, based on the same norming samples.

Selection of an achievement test battery for a school district is a very important decision. It is usually desirable to use such a battery at several grade levels and over a period of years to obtain comparable data for use in evaluating student progress. The emphases in the battery can, over a period of years, influence students and teachers in their distribution of study time.

In making their own judgments concerning the value of a test for local use, members of a test selection committee should consult the reviews of published tests in *Buros' Mental Measurements Yearbooks* and in professional journals, taking advantage of the judgment of experts to help narrow down the number of test batteries to be considered. For each test battery under consideration, committee members should examine the test itself (preferably by taking the test). Such an examination should reveal the extent to which the content is keyed to the objectives of the local course of study and should identify factors that might invalidate results for local groups (for example, disproportionate emphasis on specific facts, test so easy that it would fail to measure the best achievers adequately, and the like). The committee should also summarize the data in the manual relevant to topics considered in the summary form in Chapter 5.

Achievement Test Batteries for Both Elementary and Secondary Schools

Of the most widely used achievement test batteries, only two span the elementary and secondary school years with articulated tests measuring essentially the same pattern of learning outcomes throughout all grades. Those two batteries are the *California Achievement Tests* (CAT) and the

Sequential Tests of Educational Progress (STEP). These batteries have both been included in dual standardization programs (that is, with a companion test of scholastic aptitude, the *California Test of Mental Maturity* and the *School and College Ability Tests* respectively, being administered to the same students). The grade range for the CAT tests is grades 1 through 14; for the STEP tests, grades 4 through 14.

There are major differences between the two batteries. The CAT tests limit their coverage to the fundamental skills of reading, mathematics, and language,¹¹ while the STEP series extends the communication skills to include listening and writing skills and also includes tests in social studies and science. The STEP series is the only achievement test battery now available that includes (1) a test of listening ability and (2) an essay test, in which students write themes on selected topics, which are to be evaluated by semiobjective, clearly defined standards. Both test batteries are so designed that the user can choose to purchase and administer tests in one or more of the major subject areas.

The STEP tests are designed to measure "critical skills in application of learning." Devised with the advice and assistance of educators recommended by national professional groups (in English, mathematics, and other subject areas), these tests place minimum emphasis on memory and greater emphasis on the higher levels of cognitive abilities. The following summary of skills sampled by the science and social studies tests will help the reader to understand this emphasis on generalized outcomes, rather than upon knowledge of specifics:

Skills sampled in STEP science tests

The ability to

1. Identify and define a scientific problem.
2. Suggest, screen, and test a hypothesis.
3. Design experiments and to collect data.
4. Interpret data and draw conclusions.
5. Evaluate critically the printed and spoken word.
6. Reason quantitatively and symbolically.

Skills sampled in STEP social studies tests

Ability to

1. Read and interpret maps, charts, cartoons, pictures, diagrams, as well as the printed word.

¹¹ The California Test Bureau publishes separate tests in study skills, social studies, and science, listed in the Appendix under the titles *California Study Methods Survey* (grades 7-13) and *California Tests of Social and Related Sciences, Elementary* (grades 4-8) and *Advanced* (grades 9-12).

2. Think critically, to distinguish fact from opinion, and to recognize propaganda.
3. Assess and interpret data.
4. Apply appropriate outside information and criteria.
5. Draw valid generalizations and conclusions.¹²

This type of test design, in which students are presented with unfamiliar problems, has a few disadvantages. Fewer items of this type can be administered within an hour of testing time. Hence, with the exception of the essay test, each of the tests (in reading, language, and the like) requires 70 minutes of testing time. Another characteristic of this approach is that the tests become highly verbal tests; that is, a large percentage of the variance in scores on all tests of the series is attributable to differences in the students' verbal ability. At the fourth-grade level, scores on the STEP mathematics test correlated more highly with the verbal scores of SCAT (*School and College Ability Test*) than with the quantitative scores.¹³

The STEP series is outstanding in its carefully designed test items, its use of percentile bands¹⁴ to emphasize errors in measurement, and its aids to the teacher and students in the interpretation of test results. The CAT series also has many aids to teacher use of test results. The Scoreze booklets for the CAT tests combine the advantages of machine scoring with carbon copies of answer sheets, which clearly indicate the items each student has missed and the types of learnings tested by them. Diagnostic analyses have also been prepared for each test. Although these analyses have been criticized for giving users a misleading impression of the diagnostic value of the test, they can provide diagnostic leads. However, only a few items of each type are included; hence teachers must obtain further evidence concerning the validity of any hypotheses growing out of their study of these diagnostic clues. The use of this type of aid is considered further in Chapter 14.

Both tests provide national norms. The CAT provides grade placement norms throughout the full grade range, although their use at the high school level is of limited value. Percentile norms for each grade are also provided. The STEP tests use only percentile norms. Although grade placement norms have certain limitations, discussed in Chapter 2, they also have certain advantages for the elementary school grades (in the measurement of gains, and in the comparison of class and school results with national averages).

¹² *Cooperative Sequential Tests of Educational Progress* (Princeton, N. J.: Educational Testing Service, 1957).

¹³ Anne Anastasi, *Psychological Testing* (New York: The Macmillan Company, 1961), p. 448.

¹⁴ See illustrative profile in Chapter 5, page 164.

The provision of anticipated achievement norms constitutes a unique feature of the CAT tests. Use of AAGP norms, discussed in Chapter 14, makes it possible to compare the scores of each student with those for students in the norming sample of comparable age, grade, and scholastic aptitude and to know the approximate standard error for this type of comparison.

It is apparent that each of these two series incorporates many features designed to make it valuable for use in schools. Although a few differences have been pointed out between the two series, the choice between them, or among these and several other batteries available should be made chiefly on the basis of (1) the examination of test content in terms of its validity for the local educational program, (2) a careful study of their manuals, and (3) the reviews in the Buros Yearbooks.

Achievement Test Batteries for the Elementary and Junior High School Grades

It is easy to understand why several achievement test batteries do not include the upper secondary school years. In many school districts, city-wide achievement testing programs are conducted only in grades 1 through 8. In the higher grades, students begin to take differentiated programs; hence, the all-school testing program often becomes limited to tests designed to aid students in their choices of curricula and in post-high-school planning.

In addition to the CAT and STEP tests, four widely used achievement test batteries are available for the elementary school grades.

IOWA TESTS OF BASIC SKILLS (GRADES 3-9) Includes tests of vocabulary, reading comprehension, language skills (spelling, capitalization, punctuation, usage), and work-study skills

This test includes a test on work-study skills for even the youngest children. Its emphasis is on functional skills, rather than specific information. More subscores in language are provided than in some other tests. The battery is unusual in that tests for all areas and grade levels are included in one spiral-bound reusable test booklet. Tests for each grade are adapted specifically to that grade but utilize some of the test items for adjacent grades.

METROPOLITAN ACHIEVEMENT TESTS

Primary I (for second half of grade 1) includes tests of word knowledge (sight vocabulary), word discrimination (selecting orally presented word from set of printed words), reading comprehension, arithmetic concepts and skills.

Primary II (for grade 2) includes these tests plus spelling.

Elementary (grades 3-4) adds language and two separate tests on arithmetic (computation, and problem solving and concepts).

Intermediate (grades 5-6) has a partial battery, including the tests in the

Elementary battery (with the exception of word discrimination). The complete battery also includes science and two tests in social studies (information and study skills).

Advanced (grades 7-9) has partial and complete batteries with the same subtests listed above.

The range of content of the Metropolitan batteries for grades 5 and above is almost as comprehensive as that for the STEP tests, except that the latter includes skills in listening and writing.

SRA ACHIEVEMENT SERIES

Four batteries (grades 1-2, 2-4, 4-6, and 6-9), measuring skills and understanding in four general areas: (1) reading, (2) arithmetic, (3) language, and (4) work-study skills. All four batteries include under reading subtests on (4) work-study skills. All four batteries include under reading subtests on comprehension and vocabulary; in addition, the battery for grades 1-2 includes a verbal-pictorial association subtest, which measures the ability to comprehend isolated words, phrases, and sentences and a language perception subtest (including auditory discrimination, visual discrimination, and sight vocabulary). Language arts tests for the highest three levels provide subtest scores on capitalization and punctuation, grammatical usage and spelling. Arithmetic tests in all batteries provide subtest scores in reasoning, concepts and usage, and computation. The work-study skills tests at the two higher levels provide subtest scores in references and charts.

This test series differs from the other batteries in at least two respects.

1. Instead of assuming high-level motivation on the part of all students, the authors have presented the items in the three lower batteries in story form in order to elicit greater pupil interest. In the preliminary try-outs, student reactions to story interest were obtained and revisions made in terms of these reactions. In order to keep the tests in arithmetic and language from being influenced unduly by reading comprehension ability, special effort was made to keep the reading difficulty of these tests well below the reading level of the grade being tested.

- The story approach also represents an attempt to make the test situation less artificial and measurement of criterion behavior less indirect. For example, vocabulary items are presented in context, with the pupil selecting the meaning appropriate to the context. The arithmetic test items are the meaning appropriate to the context. The arithmetic test items are grouped around simulated situations in such a way that the student's performance can be readily interpreted in terms of failure to read problem correctly, illogical reasoning, inappropriate arithmetic procedure, or errors in computation. The story approach has resulted in greater test length, just as the approach used in the STEP series required longer tests to achieve adequate reliability.
2. The tests are designed for pupils who are achieving within the average range for their grade level or who show superior achievement for that grade level. The slow learners will be identified but not measured by the test. The authors recommend that academically retarded pupils be given the battery designed for the next lower level. Since this decision represents one solution to a difficult problem facing all authors and publishers of achievement test batteries, the discussion of this problem is quoted from the Technical Supplement for that test.

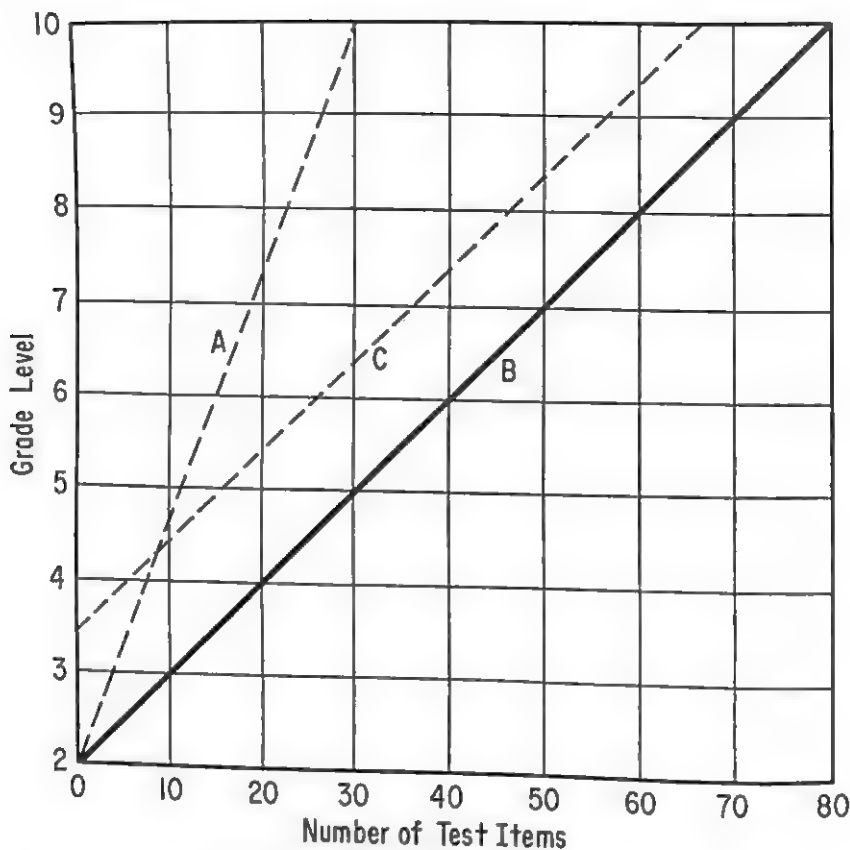


Fig. 13.2 Effects of Various Test Gradients.

Reproduced with the permission of Science Research Associates from Louis P. Thorpe, D. Welty Lefever, and Robert A. Naslund, *Technical Supplement, SRA Achievement Series* (Chicago: Science Research Associates, 1957), p. 5.

Three possible policies for the construction of the tests and the sampling of items could have been followed. Each of these procedures has advantages as well as characteristic weaknesses:

1. The test could be made to cover a wide range of abilities in a given curricular area, but with relatively few items at each level of difficulty [Figure 13.2, line A]. It would contain items appropriate for severely retarded, as well as for gifted, pupils. At least two serious limitations are likely to characterize such a test. First, the small number of items at each difficulty level will not yield dependable measures of pupil achievement. In short, there would be too few items having just the right range of difficulty for each pupil. Second, a reduced sample of test items would also result in inadequate curricular coverage at each grade level.

2. The test could be designed to contain a rich sampling of items at all

curriculum levels, appropriate both to the poorest achiever and the most skilled pupil. While an excellent measure of achievement at all levels of ability would be assured, the total testing time would rise beyond feasible limits. In the fifth grade, for example, in order to develop an adequate reading test in terms of this policy, it would be necessary to include a considerable quantity of reading material with questions suitable in difficulty for children all the way from the second grade to the tenth grade [Figure 13.2, line B]. The other levels of ability involved would require similar sampling.

3. A third policy, the one adopted for the *Series*, is to provide an adequate sample of items for each difficulty level covered, but at the same time to reduce the total range of difficulty by "lopping off" the lower end of the scale [Figure 13.2, line C]. This design is unusual for achievement tests and thus requires a brief explanation.

Each battery has been so constructed that it does not contain easy items suitable for the seriously retarded pupil to answer correctly, and only a relatively few items simple enough for the low-average learner to handle successfully. However, the upper level of each test battery has been extended sufficiently so that it overlaps the next higher battery to a considerable extent.

For example, the *SRA Achievement Series, 2-4* has a range from grade 2.5 to grade 4.9. Since this battery might be administered to a fourth grade class at the close of the school year, it contains test items difficult enough to measure the most able fourth grade pupil at that time. This means that this battery includes a considerable sampling of fifth and sixth grade test items.¹⁵

Both Hopkins¹⁶ and Sax¹⁷ have found that it is possible for students marking items at random to obtain chance scores almost at grade level on leading achievement batteries. Hence, it is highly desirable that either tests be lengthened (alternative 2) or that tests disclaim the ability to measure over such a wide range (alternative 3) and test users assume responsibility for administering less difficult tests to slow learning pupils.

STANFORD ACHIEVEMENT TEST

Primary (grades 1-3) includes two tests of reading ability (word meaning and paragraph meaning); two tests of arithmetic (arithmetic computation and arithmetic reasoning), and a spelling test.

¹⁵ Louis P. Thorpe, D. Welty Lefever, and Robert A. Naslund, *Technical Supplement, SRA Achievement Series* (Chicago: Science Research Associates, Inc., 1957), pp. 4-5.

¹⁶ Kenneth D. Hopkins, "Validity Concomitants of Various Scoring Procedures Which Attenuate the Effects of Response Sets and Chance," unpublished doctoral thesis, University of Southern California, 1961.

¹⁷ Gilbert Sax, "Theoretically Derived Chance Scores and Their Normative Equivalents on a Selected Number of Standardized Tests," *Educational and Psychological Measurement*, vol. 22 (Autumn 1962), pp. 573-576.

Elementary (grades 3–4) includes the tests listed above and adds a three-part language test (capitalization and punctuation, sentence sense, and usage).

Intermediate (grades 5–6) includes the tests listed above in its partial battery.

In the intermediate battery, complete, are also included additional tests on social studies and science.

Advanced (grades 7–9) has the same organization as the intermediate battery.

The student will recognize that this test covers essentially the same areas as the *Metropolitan Achievement Tests*, published by the same company. The Metropolitan, however, has tests of two levels of difficulty for the primary grades and includes three subtests of reading. The content of the *Stanford Achievement Tests*, in the social studies and science areas, has been criticized in the *Buros Yearbooks* and elsewhere as emphasizing unrelated factual questions, rather than the understanding of significant concepts and principles. These criticisms, however, do not apply to the 1964 edition, which is based on a content analysis of more recently published textbooks.

Achievement Test Batteries for the High School Level

In addition to the CAT and STEP series discussed earlier, there are three other achievement test batteries designed for use at the high school level. All three are designed to help predict student success in college and to help students recognize areas of weakness in their preparation for college. However, they differ markedly in their length, cost, and the types of educational outcomes emphasized.

The Iowa Tests of Educational Development (ITED) were developed to measure the student's progress in the development of broad intellectual skills. These tests emphasize understanding of what the student has learned and his ability to apply his learnings, rather than his recall of specific facts. They resemble the STEP tests in their basic approach. The STEP tests differ from the ITED in that they represent an articulated series from grades 4 through 14 and in that they include tests of listening and essay writing.

The following comparison of tests in the two series may be useful:

STEP, LEVEL 2 GRADES 10–12		ITED GRADES 9–12	
Mathematics	Test 4	Ability to do quantitative thinking	
Science	{	Test 2	General background in the natural sciences
		Test 6	Ability to interpret reading materials in natural sciences
		Test 9	Use of sources of information

STEP, LEVEL 2
GRADES 10-12ITED
GRADES 9-12

Social Studies	{	Test 1	Understanding of basic social concepts
		Test 5	Ability to interpret reading materials in social studies
		Test 9	Use of sources of information
Reading	{	Test 7	Ability to interpret literary materials (See also tests 5 and 6 above)
		Test 8	General vocabulary
Writing		Test 3	Correctness and appropriateness of expression
Essay			
Listening			

It is evident from this comparison that the ITED has more test material than STEP in the areas of social studies and science. In each of these subject areas, one can note from ITED scores whether a student seems to have a deficiency in background knowledge, ability to interpret text materials in the field, or in ability to locate and use reference materials (as reflected in test 9). On the other hand, STEP offers more scores in the important area of communication skills.

Expectancy tables have been developed so that one can predict from a student's scores at any grade level (in high school) his probable score on the tests of the College Entrance Examination Board, and his probable academic success in three types of colleges. Profiles are available that show the average ITED scores earned in high school by students majoring in eleven different college areas. Each student receives a profile leaflet entitled "Your Scores on the ITED and What They Mean," while counselors receive copies of a guide on "How To Use the Test Results."

The total score predicts college grades with unusually high predictive validity, with some validity coefficients approaching .60. This level of validity is attributable in part to the length of the test, eight hours of student testing being required unless shorter testing times (optional with the user) are employed.

The *Essential High School Content Battery* is a shorter test, which can be administered in three and a half hours. Although norms are available for students in commercial, science, general, and academic curricula, the test is best suited to college-preparatory students. The four subtests (mathematics, science, social studies, and English) cover typical content of required courses. Greater emphasis is placed on factual knowledge than in the ITED. However, items are well constructed; and the science section includes items on application of principles.

For a high school that wants to measure student achievement of im-

mediate objectives in these academic areas and use the total score as a basis for predicting college success, this test provides an economical substitute for the longer STEP or ITED batteries.

Still less time (only two hours) is required for administration of the *Cooperative General Achievement Tests* for grades 9–13. This battery includes tests in three subject areas (social studies, natural sciences, and mathematics). The subtests tend to emphasize generalized outcomes more than does the *Essential High-School Content Examination*. Each test has two parts, Part I emphasizing an understanding of terms and concepts and Part II, the ability to interpret materials and problems relevant to the field. This battery is essentially designed to measure proficiency in three areas important to future achievement in college. English is not included, probably because of the widespread use of the *Cooperative English Tests* (in reading comprehension and English expression). If the English tests are also administered, the total time is comparable to that for the *Essential High School Content Examination*.

End-of-Course Achievement Tests for the High School Level

The student will find many tests in high school subjects listed in the Appendix. Only through an examination of such tests and their manuals can he appraise their validity for his own purposes. It is desirable, however, to mention two leading series of course-oriented achievement tests. In each of these series, a common score scale is used in all tests of the series.

The Cooperative Test Service of the American Council on Education pioneered in the development of end-of-course examinations in high school subjects. Subject-matter specialists in each high school subject worked with specialists in test construction to develop a large number of tests that would be acceptable to many teachers. In 1948 this test service became the Cooperative Test Division of the Educational Testing Service.

The various revisions of the Cooperative Mathematics Tests have included a comprehensive 80-minute test for grades 7 through 9 and one-period tests in (1) elementary algebra (through quadratics), (2) intermediate algebra (quadratics and beyond), (3) plane trigonometry, (4) plane geometry, and (5) solid geometry. In 1962, new tests in arithmetic, algebra, and geometry were published. These tests were designed to reflect some of the newer emphases in mathematics, but important aspects of traditional mathematics are also measured. Additional tests in third-year algebra, trigonometry, analytic geometry, and calculus are to be published in 1964.

The *Cooperative Science Tests* include an 80-minute test for grades 7 through 9 and also one-period tests in (1) general science, (2) biology, (3) chemistry, and (4) physics. A special series of unit tests has also

been prepared for those schools using the physics course developed by the Physical Science Study Committee. The Cooperative Test Division is engaged in a major revision of its course-oriented tests in science, which are scheduled for publication in 1964.

The *Cooperative Social Studies Tests* include an 80-minute test for grades 7 through 9 and also one-period tests in American history, American government, ancient history, modern European history, and world history. In the revised series, scheduled for publication in 1965, both junior high school and senior high school tests in American history will be available, as well as revised tests in American government, world history, and modern European history. New tests in civics and problems of democracy are also being developed.

The *Cooperative Foreign Language Tests* now include one-period tests in French (elementary and advanced), Spanish (elementary and advanced), and Latin (elementary and advanced). A listening comprehension test in French (with tape-recorded selections) is also available. In 1964 the Cooperative Test Division will publish a series of new tests, which are being developed in cooperation with the Modern Language Association and the United States Office of Education. This comprehensive testing program will cover

1. Five languages (French, German, Spanish, Italian, and Russian).
2. Four skills in each language (reading, writing, listening, and speaking).
3. Two levels in each skill (beginning and intermediate).
4. Two equivalent forms for each level.

Multiplication of the numbers listed above indicates that this program involves 80 new tests. This series includes the first standardized tests ever developed of ability to speak foreign languages. Obviously this program represents a tremendous advance in measurement in this subject area, an advance that could not have been made without financial subsidy and the cooperation of many professional workers in both the subject fields and the field of professional test construction.

Fortunately for test users, another series of high school achievement tests is also available. We use the term "fortunately" because of our conviction that the test user should be able to choose among available achievement tests in terms of their content validity for his purposes. In fact, the teacher will find listed in the Appendix other published tests that may serve his needs better than tests from either of these series.

The *Evaluation and Adjustment Series*, published by Harcourt, Brace & World, Inc., is the other leading series of high school achievement tests. This series includes more than 20 subject tests, which are listed in Appendix A in the sections on language, mathematics, reading, science, social studies, study skills, and miscellaneous.

This series offers some tests (such as those in civics, health knowledge, and psychology) that have no parallel in the Cooperative series. On the other hand, the Cooperative series includes tests in foreign language and in a few other subjects (for example, solid geometry, ancient history, and modern European history) that have no parallel tests in this series.

Two features of the *Evaluation and Adjustment Series* should be emphasized. One is the availability of *item norms*, which are of great value in group diagnosis (as illustrated in Tables 13.1 and 13.2). An additional advantage is that individual teachers can have students omit certain questions, or they can combine selected questions from two forms, as suggested in Table 13.3. By obtaining the average value of item norms for questions used, a teacher can still compare his class average with average student achievement on these items by the norming sample.

Table 13.3

Hypothetical Example of the Use of a Standardized Test as A "Final Examination"^a for Teachers' Classes That Have Given Somewhat Different Emphases to Various Aspects of the Subject

-
1. Each teacher of American history in the school district independently reviewed all items on the *two forms* of the *Crary American History Test*. On accompanying answer sheets, he rated *each item* according to the following scale:
 A—Essential concept or item of information; should be learned by every student
 B—Of major importance
 C—Fairly significant; usually covered
 D—Comparatively unimportant, or inappropriate for this grade level
 E—Inconsequential, trivial
 2. Of the 180 items on the two forms, 95 were given either an A or B rating by 10 of the district's 12 teachers of American history. These items constituted an "anchor test."
 3. Each teacher selected an additional 30–40 items that he wanted to have included in his final examination. The remaining 45–55 questions (which differed from teacher to teacher) could be deleted from answer sheets so that students would not have to spend time on them. However, if desired, all questions could be administered to students with the understanding that any question not considered in their text or class discussions would be deleted from their final examination.^b
 4. When the tests had been administered, all tests for the school district were scored on the anchor test key. Then the tests for each teacher were scored for the additional questions he wished to have included on his end-of-course examination.
 5. Local stanine norms for the anchor test were established by graphing the frequency distribution on the Otis Normal Percentile Chart.
 6. An "expected median raw score" for each teacher's "final examination" was obtained by averaging the "percentage right" values for all test questions included in his examination. These item values were obtained from the test manual. Such an expected median score, of course, made no allowance for differences between local classes and the norming population, with respect to average intelligence or

Table 13.3 (Continued)

Hypothetical Example of the Use of a Standardized Test as A "Final Examination"^a for Teachers' Classes That Have Given Somewhat Different Emphases to Various Aspects of the Subject

reading achievement. Such allowance, however, could readily be made on the basis of data provided under 7a below.

7. Each teacher received a report for each of his classes, which gave the following information:
 - a. The distribution of stanine scores for each class on a reading comprehension test that had been given early in the school year. This information provided a crude basis for judging how well the class might reasonably be expected to achieve in a course that required considerable reading.
 - b. The distribution of stanine scores^c for each class on the anchor test (composed of questions to which local teachers had assigned A or B ratings).
 - c. An item count (of the type shown on page 479) for all the questions in his "final examination." These data for all items could easily be compared with data in the test manual for the norm sample.^d
 - d. A list of students' names, with two scores for each student:
 - (1) a stanine score on the "anchor test"
 - (2) a raw score on the teacher's "final examination"

^a The term "final examination" is put in quotation marks because it is assumed that each teacher would supplement the standardized test by questions he (or he and his school colleagues) had devised, which measured outcomes that did not seem to be adequately measured by the standardized test. The teacher could combine the scores from the standardized test (or portion thereof) with the scores from his own teacher-made test, weighting the scores in any proportion which seemed best to him.

^b Under this plan, each student had to use two answer sheets, which represented an additional four cents per student for testing supplies.

^c The distributions of stanine scores on the reading and history tests for the school district were not needed as a basis for comparison. By comparing each class distribution with the standard percentages (see Chapter 2, page 41), the teacher could see how each class compared with the school district in achievement on the anchor test and on ability to read assignments. In other words, comparison with school district figures could easily be made by the teacher but were not forced upon him.

^d Local curriculum research can be facilitated by comparing school-district performance on each item with comparable data for the norm sample. For an illustrative study, utilizing this approach, see Tables 13.1 and 13.2 in this textbook and also "Curricular and Instructional Implications of Test Results," *Test Service Bulletin* No. 75 (New York: Harcourt, Brace & World, Inc., n.d.).

The second unusual feature is explained in Chapter 2. By equating the average standard score on a physics test to the average IQ of physics students, and making similar adjustments for other high school subjects, the publishers have developed a set of standard scores that is comparable from subject to subject. These norms make intraindividual comparisons

possible and facilitate the comparison of the achievement of individuals or groups with their expected achievement, as shown in Chapter 14.

The first innovation is of great value to high school teachers who want to take advantage of the item-writing skills of specialists and yet avoid tailoring their instruction to any specific test. The second innovation is of value to teachers in detecting students whose general scholastic aptitude should enable them to achieve at a higher level in a subject. It is also of special value to counselors who can obtain a better picture of a student's relative achievement in different subject fields than can be provided by comparing his grades in these subjects.

INTERPRETATION OF DATA FROM ACHIEVEMENT TESTING PROGRAMS

In comparing achievement test results for different pupils, classes, or schools, allowance must be made for differences in scholastic aptitude. The use of Anticipated Achievement Grade Placements has already been discussed as an approach that takes into account not only the student's scholastic aptitude but his age and grade. Although AAGP's are available only for the CAT series, other publishers have prepared devices similar to that presented in Table 13.4 to assist test users in modifying their expectations for students or groups whose IQ's are considerably above or below average. Other characteristics of the student population, such as the level of vocabulary and language usage in neighborhoods of low socioeconomic status, or the bilingual background of children, must also be considered in making interschool comparisons.

Most school districts administer a selected achievement test battery at several different grade levels as a basis for ascertaining whether individual students and groups have made at least a year's gain during a calendar year. Differences among classes in average scholastic aptitude must also be considered in interpreting gains.

When we attempt to make allowance for scholastic aptitude, however, another disconcerting problem, known as the regression effect, complicates the problem. A review of the diagram in Figure 3.2 will remind the reader that scores on the predicted variable tend to regress toward the mean of the group. To use less technical language, the regression effect means that students who rank high on a scholastic aptitude test are not likely to rank quite as high on a test of achievement; while those who rank low on the aptitude test are not likely to rank quite as low on the achievement test.

Durost and Prescott have illustrated this regression effect in their discussion of expected stanine scores on achievement tests for students who

Table 13.4
METROPOLITAN ACHIEVEMENT TESTS (Grades 3-4). Expected Deviations from
Grade Norm for Pupils at Successive Levels of Pintner-Durost IQ (Scale 2).

INTELLIGENCE QUOTIENT	WORD KNOWLEDGE	WORD DISCRIMI- NATION	READING	SPELLING	LANGUAGE A & B	ARITHMETIC COMPUTATION	ARITHMETIC PROB. SOLVING AND CONCEPTS
125-129	+2.1	+1.7	+2.4	+2.0	+2.0	+6	+1.2
120-124	+1.7	+1.4	+1.8	+1.6	+1.5	+4	+1.0
115-119	+1.2	+1.0	+1.2	+1.1	+1.0	+3	+7
110-114	+8	+7	+7	+7	+6	+2	+4
105-109	+3	+3	+2	+4	+3	+1	+2
100-104	-2	-1	-2	0	-1	-1	-1
95-99	-6	-5	-5	-3	-4	-2	-4
90-94	-9	-8	-8	-7	-7	-3	-6
85-89	-1.1	-1.0	-1.0	-1.0	-9	-4	-8
80-84	-1.3	-1.2	-1.3	-1.3	-1.1	-5	-1.0
75-79	-1.4	-1.4	-1.5	-1.6	-1.3	-6	-1.2
Correlation between IQ and Subtest	.84	.82	.81	.73	.71	.56	.71

Source: Metropolitan Achievement Tests: Expected Deviations from
 Grade Norms for Pupils at Successive Levels of Pintner IQ, Test Data
 Report No. 18 (New York: Harcourt, Brace & World, Inc., 1961).

Publisher's Note: In interpreting individual performances, consider-
 able leeway should be allowed since both the pupil's IQ and his
 Metropolitan grade equivalent are subject to some measurement
 error. In general the higher the correlation [in the last row], the
 more reliable the tabled deviation values.

are above average or below average in intelligence.¹⁸ For example, if the correlation between an intelligence and achievement test were .70, the most probable achievement test scores for individuals at each intelligence or capacity stanine would be as follows:

STANINE ON TEST OF INTELLIGENCE OR SCHOLASTIC APTITUDE	MOST PROBABLE STANINE SCORE ON ACHIEVEMENT TEST
9	7.8
8	7.1
7	6.4
6	5.7
5	5.0
4	4.3
3	3.6
2	2.9
1	2.2

When we examine this table, we can see that a student with an intelligence stanine of 9 and an achievement stanine of 8 cannot be labeled an "under-achiever."¹⁹ It should not be difficult for the student to see that the same principle is involved when a superintendent compares the achievement data for schools. It is difficult for a high-ability school to show as high an average score in achievement as in scholastic aptitude.

Findley has made a suggestion for interpreting data for schools that would not only take into account this regression effect but many other factors that cause a group's achievement to be initially lower or higher, such as the low level of motivation among children who are academically retarded or come from homes of low cultural background. He suggests that we compare the gains of schools *with similar initial median scores* in a subject field.

Factors of native ability, subcultural motivation and general instructional effectiveness to date have combined to produce initial medians at the grade

¹⁸ Walter N. Durost and George A. Prescott, *Essentials of Measurement for Teachers* (New York: Harcourt, Brace & World, Inc., 1962), p. 87.

¹⁹ We can easily check the results in this little table by translating stanine scores into deviations from the mean stanine of 5, and using the simple standard-score regression equation given on page 79. (This equation can be used because stanines are linear transformations of z-scores.) Since a stanine score of 9 is a deviation score of 4, we can substitute this value of 4 and our r of .70 in the standard-score regression equation to obtain: Predicted stanine deviation (for achievement) $= (.70)(4) = 2.8$. This deviation value is added to the mean of 5 to obtain the first value in the table. All others can be checked in the same way, or the regression effects occurring when r has a smaller or larger value can be obtained.

level in question. It may be fairly assumed that the same factors will continue to operate in similar measure thereafter except as growth data show the classes in a school to have improved more than others *with a similar start*. [Italics added]²⁰

Table 13.5

Three-Year Gain Scores in Arithmetic (1956-1959) for Schools with Various Median Scores When Tested at Grade 3.1 in Fall 1956

THREE-YEAR GAINS (IN YEARS)	Initial Median Grade Placements			
	1.5-1.9	2.0-2.4	2.5-2.9	3.0-3.4
4.0-4.4			1	1
3.5-3.9		1	8	6
3.0-3.4		3	20	5
2.5-2.9	1	1	15	1
2.0-2.4	3	3	4	
1.5-1.9	7	7	1	
1.0-1.4	3	2		
Median gain	1.7	1.9	3.1	3.5

Source: Adapted from Warren G. Findley, "Gains vs. Status Scores as Evidence of Effectiveness of Instruction," *The 19th Yearbook, National Council on Measurement in Education* (Ames, Iowa: The Council, 1962), p. 19.

In Table 13.5, the median three-year gains in arithmetic are shown for schools with a similar start. All data are for schools initially tested in the fall of grade 3. It is difficult to conceive a fairer basis for interpreting data for schools. Such a procedure would allow each school to see how it compared with other schools with a similar start, without revealing the specific identities of the schools being compared. For example, in Table 13.5, a school with an initial median arithmetic score in the 1.5-1.9 group would know that their gain was above average for similar schools if they had progressed at least two years during the three-year period studied. It would probably not be difficult for publishers to provide such data on one-year or two-year gains for a representative sampling of schools. Selective immigration into, or outmigration from, the school neighborhood would still have to be taken into account in interpreting data concerning gains for individual schools.

Tables 13.1 and 13.2 illustrate the way in which item norms can be

²⁰ Warren G. Findley, "Gain vs. Status Scores as Evidence of Effectiveness of Instruction," *19th Yearbook, National Council on Measurement in Education*, (Ames, Iowa: The Council, 1962), pp. 17-20.

used to discover how the local instructional program differs in emphasis from the schools used in norming the test. It is doubtful, however, whether differences should be labeled by the terms "strengths" and "weaknesses" (as in Table 13.2). The differences merely represent reliable differences between local performance and "national" performance. It is only when these measurement data are screened against *intended* local emphases that an evaluation (or value judgment) can be made.

It is desirable in any such curriculum study to have teachers agree *in advance* concerning (1) test items on which they hope to exceed national item norms, because of special emphasis in their local program and (2) test items on which they could accept lower-than-average results without apology because they have intentionally minimized these understandings or skills in order to allow proportionally more instructional time for others which they deem more important.

It is also essential, in interpreting results of standardized achievement tests, to recognize that they measure only a portion of all educational outcomes. As will be emphasized in Chapter 15, evidence should be obtained concerning all important educational outcomes by the most dependable techniques available. If plans are not made for a comprehensive evaluation program, there is a real danger that the outcomes measured by standardized achievement tests will be overvalued.

LARGE-SCALE TESTING PROGRAMS

Widespread concern has developed about the impact on students and school programs of the proliferation of large-scale testing programs.²¹ College-bound high school students, for example, may take college admission tests, tests for national scholarship programs, tests administered by their own school districts, and sometimes additional tests given by a state department as a basis for "quality control."

One concern has to do with the amount of time devoted to such programs. Certainly, every effort should be made to use students' time to advantage and to avoid needless duplication of testing efforts. However, many national testing programs, such as the CEEB tests, are administered on Saturday; the testing time does not interfere with the regular school program. Moreover, these tests are *optional* with the student.

Another concern has to do with student tension concerning examinations. Tension can be reduced to some degree when materials are presented

²¹ An example of criticism in a news magazine is "A Rash of Testing in the Schools. Is It Being Overdone?" *United States News and World Report*, vol. 46 (June 15, 1959), pp. 44-46.

to students in advance of testing that explain the purposes of a testing program, give illustrative examples of test items, and indicate how the test results will be reported to them and their schools.²²

Another concern is that large-scale testing programs may have adverse effects on the instructional program if teachers focus their attention on coaching for the test. If city-wide or state-wide programs of testing focus narrowly on specific learnings, and comparisons are made among classes and schools, this type of deleterious effect is almost inevitable. During the early emphasis on survey testing such effects occurred and caused many teachers, as well as many leaders in curriculum and supervision, to feel antagonistic toward any type of achievement testing that did not originate with the teacher. Certainly tests used on a large-scale basis should stress the most important goals of education and allow for flexibility and variety in the means by which these goals are achieved.

Another hazard of large-scale testing programs is that of using the scores routinely in making decisions about people.²³ In the selection, placement, and guidance functions of the school, teachers and counselors must use all the relevant data that we have about students and, when the student has sufficient maturity, help him in the interpretation of all the data relevant to choices he wishes to make.

All of these hazards seem to be ones that can be avoided by the proper planning of testing programs and appropriate use of test results.

SUMMARY STATEMENT

Standardized achievement tests were widely used only after World War I. Studies revealing great subjectivity and unreliability in teachers' marks stimulated the construction of objective, standardized tests in many curriculum areas. Comprehensive batteries of achievement tests were developed with which students' comparative achievements in such areas as reading, arithmetic, spelling, and language usage could be determined and could be compared with the achievement of other students in their own grade.

The first standardized objective tests were used extensively in city and state surveys. Publication of survey results led to recognition of the wide range of individual differences among students of a single grade level and aroused the teaching profession to the need for adapting instruction to individual differences. The uncritical use of survey test results, however, led in some instances to an overemphasis on the measured outcomes of instruction and a consequent

²² An excellent example of such orientation material is *A Description of the College Board Achievement Tests* (Princeton, N. J.: Educational Testing Service, 1962).

²³ A helpful bulletin on this topic is William C. Daly, "Test Scores: Fragments of a Picture," *Test Service Notebook* No. 24 (New York: Harcourt, Brace & World, Inc., 1959).

narrowing of the educational program. Curriculum changes and the child-study approach led to the development of a broader and richer conception of evaluation activities, including the appraisal of understandings, appreciations, attitudes, and interests.

Standardized achievement tests are now used in most school systems to serve a variety of purposes. The achievement tests considered for use must be appraised in terms of their value for each of the major purposes for which they will be used.

Each of the most widely used achievement test batteries was briefly described with special attention to organization of test content and any unique features that have been developed. Two series of end-of-course achievement tests for use in high school subjects were also described. Recommendations were made concerning the ways in which standardized end-of-course examinations might be used with due respect for the latitude professional teachers should have in adapting instructional programs to the needs of specific classes. The values and the problems associated with large-scale testing programs were briefly considered.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

- CASSELL, RUSSELL N., AND EDWARD J. STANCIK, "Factorial Content of the Iowa Tests of Educational Development and Other Tests," *Journal of Experimental Education*, vol. 29 (December 1960), pp. 193-196.
- CROOK, FRANCES E., "The Classroom Teacher and Standardized Tests," *Teachers College Record*, vol. 58 (December 1956), pp. 159-168.
- DIEDERICH, PAUL B., "Pitfalls in the Measurement of Gains in Achievement," *School Review*, vol. 64 (February 1956), pp. 59-63.
- EBEL, ROBERT L., AND F. M. RAUBINGER, "A Nationwide Testing Program—Opinions Differ," *National Education Association Journal*, vol. 48 (November 1959), pp. 28-29.
- FERRIS, FREDERICK L., JR., "Testing in the New Curriculums: Numerology, 'Tyranny,' or Common Sense?," *The School Review*, vol. 70 (Spring 1962), pp. 112-131.
- FINDLEY, WARREN G., "The Ultimate Goals of Education," *School Review*, vol. 64 (January 1956), pp. 10-17.
- KATZ, MARTIN R., *Selecting an Achievement Test: Principles and Procedures*. Evaluation and Advisory Service Series, No. 3. Princeton, N.J.: Educational Testing Service, 1958. Available on request.
- SEASHORE, HAROLD, AND J. E. DOBBIN, "How Can the Results of a Testing Program Be Used Most Effectively?" *Bulletin of the National Association of Secondary School Principals*, vol. 42 (April 1958), pp. 64-68.
- TRAXLER, ARTHUR E., "Use of Results of Large-Scale Testing Programs in Instruction and Guidance," *Journal of Educational Research*, vol. 54 (October 1960), pp. 59-62.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. When did standardized achievement tests begin to flourish in America? What factors gave impetus to the movement? What factors led to a reaction against standardized testing by many supervisors and teachers?

2. Interview the director of testing of a school system that has an organized achievement testing program. Describe the program in terms of the types of tests used, the frequency of their administration, and the specific uses of test results.

3. List a number of reasons why standardized tests are used less frequently in social studies than in the basic skills. What other types of evaluation techniques are used to appraise student growth toward the goals of the social studies?

4. Compare and evaluate the subtests on social studies or science in two or more achievement test batteries.

5. Describe and evaluate two or more standardized achievement tests in your major subject field. Consult the reviews in the *Buros' Mental Measurements Yearbooks*.

6. Discuss the values and limitations of such item analyses as are presented in Tables 13.1 and 13.2 in helping a teacher to identify points of possible over-emphasis or underemphasis in his instructional program.

7. Discuss the relative advantages and disadvantages of the three possible policies with respect to the range of difficulty of items in a standardized test, as discussed in this chapter.

If individualized instruction is to attain maximum effectiveness, it must be based on educational diagnosis. Although the term "diagnostic study" implies a detailed study of the learning difficulties of an individual student, the more general term "educational diagnosis" includes within its scope all activities in measurement and interpretation that help to identify growth lags and their causal factors for individuals or for class groups.

Five levels of diagnosis have been identified by Ross and Stanley. Of these, three are emphasized in this chapter: (1) Who are the pupils having trouble? (2) Where are the errors located? and (3) Why did the errors occur? They list two still higher levels in the process of educational diagnosis: (4) What remedies are suggested? and (5) How can the errors be prevented.¹ The first four levels are concerned with *corrective* diagnosis; the fifth with *preventive* diagnosis—the discovery and modification of preventable factors that are within the control of the school.

MEASUREMENT AS BASIC TO EDUCATIONAL DIAGNOSIS AND INDIVIDUALIZED INSTRUCTION

The following tenets indicate the author's point of view regarding the importance of educational diagnosis and individualized instruction, as well as the place of measurement in these aspects of education. These tenets serve not only as an introduction to our study of diagnosis but summarize the principles implicit in Part III on the use of measurement in the improvement of instruction.

¹ C. C. Ross and Julian C. Stanley, *Measurement in Today's Schools*, third ed. (Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1954), p. 332.

Interindividual Differences in Ability and Achievement Pose Crucial Problems for the Teacher

A teacher soon recognizes that the students in his class differ widely in most of the characteristics related to learning. Some have little or no interest in the instructional activities of the class, whereas others have an intense desire to learn. Some have a very meager vocabulary and a limited experiential background, whereas others have had rich cultural experiences in the home, in the community, and through extensive travel. Some have physical handicaps, poor health, impaired vision and hearing, which hinder success in learning; others are vitally alert and well equipped to attain easy success, both within and outside the classroom. The teacher is constantly aware that some students learn easily; whereas others find the work difficult, respond slowly to instruction, and require much practice.

Intraindividual Differences Must Also Be Studied

Knowledge of intraindividual differences is usually even more significant for teaching than knowledge of *interindividual* differences. A student, for example, may be below average for his grade in reading, markedly above average in arithmetic computation, and approximately at grade average in other subjects. It is not unusual for a student to show a range of three to four years in his achievement in the various subtests of an achievement battery. In addition to individual differences in achievement, there are often highly significant differences in personality traits, cultural background, and interests which affect students' educational needs.

Studying and Understanding Students Are Essential to Good Teaching

Materials and techniques of instruction, if they are to be effective, must be geared to the learning needs of students. These needs are related to students' aptitudes, interests, physical handicaps, and adjustment problems, as well as to their specific learning difficulties. Early recognition of a problem reduces the need for corrective work and prevents the development of more serious problems.

Teachers must become sensitive, therefore, to student behavior that reveals lack of interest or feelings of discouragement and frustration. Such behavior must be recognized as an expression of need—as symptomatic of problems that must be identified before the student can make normal progress in academic work. Understanding the causal factors contributing to a student's problem is essential to planning a corrective program. Studying all students in order to understand their problems is a fundamental aspect of effective teaching.

Standardized Tests Are Important Tools to Aid Teachers in Understanding Students

For students who are retarded academically, the instructional program usually requires some modification. Results of standardized tests assist in determining the optimal difficulty of instructional materials to be used. Such students frequently have failed to master all the essential skills taught at preceding grade levels. Test results can aid in analyzing their difficulties and in ascertaining which skills, information, and concepts they have failed to acquire.

Standardized tests can provide data that are more objective, more precise, and more analytical than those that result from unaided teacher judgment. Although teachers can usually identify students who are good and poor achievers, they are usually unable to identify the *nature* of the strengths and weaknesses in individuals *within* a subject field.²

Standardized Tests Need To Be Supplemented by Other Techniques of Studying Students

In the diagnosis of the causal factors underlying poor academic achievement, as well as in the study of such problems as tenseness or predelinquent behavior, the results of standardized tests must be supplemented by extensive data relating to the student's attitudes and feelings.

The informal methods of child study, discussed in Chapter 8, can aid the teacher in developing a more adequate understanding of each student. These techniques involve direct study of the spontaneous behavior of the student in the many interrelationships of his everyday life. The teacher, however, must be aware of the limitations as well as the advantages of these subjective methods. Adequate training in, and discriminating use of, informal methods can make the teacher increasingly aware of recurring patterns of student behavior and their significance as an expression of the student's needs and problems.

Data Derived from Studying Students Must Be Interpreted Objectively

In recording and summarizing data obtained through the more informal methods, the teacher must make certain that his interpretations are free from prejudice and personal bias. The trained observer knows how to use

² John C. Flanagan, "The Critical Incident Technique in the Study of Individuals," *Modern Educational Problems*, Report of the Seventeenth Educational Conference (Washington, D.C.: American Council on Education, October 30-31, 1952).

the scientific method in organizing and interpreting data. He makes tentative hypotheses about possible causes of recurring behavior, and he examines data that might support or refute these hypotheses. He recognizes that conclusions are justified only when evidence from different methods converge or tend to support the same hypothesis.

Measurement, Evaluation, and Individualized Instruction Are Interrelated Aspects of Effective Teaching

Evaluative judgments that are not based on reliable measurement, or the convergence of data obtained by different informal methods, may be incorrect and misleading. The significance of small differences in test scores may be overestimated, or wishful thinking may replace objective appraisal. On the other hand, measurement that does not lead to evaluative judgments and to wiser decisions is of little value.

Measurement of the strengths and weaknesses of a class, and of individual students, is essential as a basis for planning each step in an instructional program. Such a study reveals the need for providing instructional materials that are differentiated with respect to difficulty and content; it also points out specific needs for corrective teaching. Instruction that is not individualized to meet the needs revealed by measurement tends to be ineffective and frustrating to teacher and students alike. Measurement, evaluation, and individualized instruction are therefore interrelated components of effective teaching.

The School Must Accept Responsibility for Developing Every Student to the Maximum of His Potentialities

To teachers who accept responsibility for helping *all* students develop to the limit of their potentialities, differences among individuals (and among traits in the same individual) constitute a challenge. Through study and experience, these teachers gradually develop the understandings and techniques necessary for studying the needs of students and planning individualized instructional activities. Group instruction is utilized for subgroups within the class that are reasonably homogeneous in their ability to profit by such experiences. Individualized instruction is provided for students who are deficient in the skills; and special classes or other modifications in program are provided for those who show serious growth lags. The effectiveness of programmed instruction with retarded and slow-learning students has demonstrated the need for greater individualization of instruction.

At the secondary school level, the responsibility for individualizing the educational program is shared by guidance workers and teachers. The guidance workers help the student to select the courses, extracurricular activities, and work experiences that will be most valuable to him. Even if the counseling program is adequate, the classroom teachers must still teach fairly heterogeneous groups of students with a wide variety of interests and needs. The high school teacher's greater student-contact load, and the shorter length of time he has students in class, make individualization of instruction more difficult at the secondary school level. The secondary school teacher, however, can use plans and materials that capitalize on the greater maturity of the high school student and his increased ability to direct his own activities.

A small growth lag that is corrected early will not develop into a critical disability in later grades, when deficiencies in learning may be complicated by attitudes of defeatism and negativism on the part of the student, and when such deficiencies may damage the student's relationships with parents and classmates. Early identification of learning difficulties, followed by individualized instruction, constitutes economical and effective preventive work and reduces the need for later remedial work.

LEVELS OF DIAGNOSIS

Faced with the obvious fact that he cannot do comprehensive diagnostic work with each student, the teacher is rightly concerned with determining the level of diagnosis he should attempt in specific situations.

The term "level of diagnosis" cannot be defined precisely. An illustration, however, may help to clarify the concept. A teacher may administer an achievement test battery and, on the basis of the results, note that a student's difficulty is in arithmetic computation, rather than in arithmetic reasoning, reading, language, or other school subjects. The diagnosis may be carried to another level by determining in which of the arithmetic processes he is weak (for example, division). Diagnosis may be carried to a still more specific degree by discovering that the student does not know how to estimate trial divisors, although he does have an adequate familiarity with the division combinations.

No rule can be established as to the level of diagnosis that is appropriate in a specific situation. In general, it may be said, however, that a satisfactory level of diagnosis has been reached when the teacher has gained sufficient insight into the nature of the student's problem to enable him to plan appropriate corrective instruction. This will be determined in large part by the complexity of the individual problem.

As Tyler has said, "A satisfactory diagnosis should be as specific as the

desired outcomes permit and as the possibility of localization of symptoms allow, so long as the diagnosis is practicable. It need not be carried farther than is appropriate for the remedial program provided."³

STEPS IN EDUCATIONAL DIAGNOSIS

The essential steps in educational diagnosis are (1) identifying the students who are having trouble, (2) locating the errors or learning difficulties, and (3) discovering the causal factors. In the following discussion of these three steps in educational diagnosis, the illustrative material will be concerned chiefly with diagnosis in the basic skills. This section is followed by a summary of suggestions on group diagnosis and other procedures that are practicable for teachers of high school content subjects.

Identifying the Students Who Need Help

The inclusion of this step is based on a recognition of the realities of the typical teaching situation. *All* students can profit from individualized help, given on the basis of educational diagnosis. In the typical classroom, however, individualized corrective instruction can be given to only a few students.⁴ Hence, the teacher's first step is to identify the students in each subject area for whom diagnosis and corrective instruction are imperative in order for them to participate in group instruction with profit.

Although students who are in the greatest need of corrective instruction can be identified by scanning achievement test profiles, data on scholastic aptitude, and other relevant data, the following procedures will be found to be systematic and objective:

1. Plot each student's test (or subtest) score and some measure of his expected achievement on a two-variable chart, similar to Figures 14.1 or 14.2.
2. Draw a "staircase" diagonal from the lower-left- to the upper-right-hand corner of the chart, which includes at each level the cell or square for students whose converted scores on the intelligence and achievement tests correspond (fall in the same interval), as well as the adjoining cells for those whose scores differ by only one interval.
3. Note the names of all students whose scores are outside the "staircase" diagonal. Those students whose tallies are to the left of the staircase appear

³ Ralph W. Tyler, "Characteristics of a Satisfactory Diagnosis," *Educational Diagnosis*, 34th Yearbook, National Society for the Study of Education (Chicago: University of Chicago Press, 1935), p. 106.

⁴ If carefully selected programmed materials have been made available, individualized instruction for all may be possible.

to be "underachieving"; those to the right appear to be "overachieving." These two terms are tentative classifications, applied to students whose achievement scores are significantly lower, or higher, than measures of their scholastic aptitude.

4. For all students who appear to be overachieving, check to see whether the student is overachieving in other achievement tests. If he is, examine data on previous scholastic aptitude tests on his cumulative record. The most recent test of scholastic aptitude may have given him a lower IQ than he obtained on previous tests. If he is a new student, with no previous intelligence test data available, a retest should probably be given.
5. For all students who appear to be underachieving, further diagnostic work should be undertaken. More information on these individuals should be obtained as a basis for developing (a) hypotheses regarding the reasons for their low achievement and (b) suggestions for helping them toward more adequate achievement. Diagnostic study and corrective instruction are especially important if these students are below national norm (which represents society's standards for that grade level) or if their level of achievement is likely to handicap them in the pursuit of their educational and vocational goals. On the other hand, a brilliant student planning to major in literature would not be considered as needing diagnostic study if his stanine in intelligence were 8 or 9, and his score in mathematics or science were 2 stanines lower.

USING LOCALLY DEVELOPED NORMS ON A TWO-VARIABLE CHART Figure 14.1 illustrates a type of two-variable chart that can be used with any standardized test that has stanine scores, for example, the *Metropolitan Achievement Tests*. Stanine scores represent equal units of $\frac{1}{2}$ SD; moreover, stanine norms are easily developed from local data.

Another unusual feature of Figure 14.1 is that data for a single class are superimposed on a two-way table for a *school*. The teacher has written in the names of the students in his class on a mimeographed table providing data for the school; such a table could be provided for a school district or larger unit. In this way, the teacher can see how many of his students are working higher or lower than expectancy (in terms of scholastic aptitude); and he can also compare the distribution of scores for his class with those for the school or school district.

For example, in Figure 14.1, only 2, or 6 percent, of the students in Miss Lee's class appear to be overachievers, as compared with 22 percent in the school. On the other hand, 48 percent of Miss Lee's class appear to be underachievers, as compared with 21 percent in the school.

Without further information, we do not know whether Miss Lee's students were grouped into a low-achiever class, whether the students have made inadequate progress in eighth grade, or both. We can easily determine that the median stanine on IQ is 4, while the median in arithmetic is one full point lower. We would especially like to have more information about the seven students for whom the difference between achievement and expectancy is very large, that is, three or more stanines.

Stanines on Local Arithmetic Test										
	1	2	3	4	5	6	7	8	9	TOTAL
9						1		2	2	5
8				WALTER		PHIL				
				1		2	3	1	1	8
7		FRED		LOIS	NINA, LOUISE	1				
		1	1	2	3	2	2	3	1	15
6		TOM	ROSE, HARRY	WAYNE, TED						
		1	2	4	5	4	1	2	1	20
5		PETER	BOB, ROY	SALLY, RITA						
		1	3	6	8	4	2			24
4		AMY, JANE	KARL, ANN	SAM,JO DONALD						
		3	2	5	4	3	3			20
3		KIM	MARY	SUE, SARA	JOHN, JIM					
		1	3	2	3	4	2			15
2	CAROL	MAY								
	1	1	2	1	2	1				8
1	MAMIE	ANNA								
	2	1	1				1			5
TOTAL	3	9	14	21	25	21	14	8	5	120

Fig. 14.1 Names of 33 Pupils in Miss Lee's Eighth-grade Arithmetic Class, Entered on a Two-Variable Chart for All Eighth-graders in Central High School.

This chart is designed to show the relationship between stanine scores on the Pintner General Ability Test and stanine scores on the local arithmetic test. (The r between these two tests is approximately .52.) All cases within the diagonal "staircase" have the same stanine scores in both the intelligence and arithmetic tests, or their stanines in these two tests differ by only one score. For all cases to the right of the diagonal "staircase" (in this case 23 percent of Central High School students), the student's arithmetic stanine exceeded his PGAT stanine by two or more points. For all cases to the left (in this case 21 percent), the student's arithmetic stanine was two or more points below his PGAT stanine.

USING THE AAGP AS A MEASURE OF EXPECTANCY In studying the results for subtests of the *California Achievement Tests*, an especially designed measure of expectancy (the Anticipated Achievement Grade Placement) can be used. One can compare each student's achievement with his AAGP, that is, the average grade placement earned on that subtest by students of his grade, age, and mental age. Table 14.1 lists reading test results for 50 students and indicates the difference between each stu-

dent's reading comprehension GP and his AAGP. Those differences smaller than one $SE_{\text{difference}}$ are placed in parentheses. Asterisks are used to indicate the level of assurance with which we can interpret other differences as being greater than chance differences, or unlikely to be reversed on retesting.

Table 14.1

Anticipated Achievement Grade Placements and Grade Placements in Reading Comprehension^a for a Class of Ninth-Grade Students

STUDENT	ANTICIPATED ACHIEVEMENT GRADE PLACEMENT ^b	TOTAL READING GRADE PLACEMENT	DIFFERENCE ^c
1. Albert	10.1	8.9	-1.2
2. Alfred	9.8	7.3	-2.5**
3. Adele	8.8	5.0	-3.8**
4. Arnold	11.3	9.1	-2.2**
5. Arleen	8.6	8.5	(-0.1)
6. Audrey	11.2	10.0	-1.2
7. Brian	10.2	10.3	(0.1)
8. Byron	9.3	8.9	(-0.4)
9. Carol	6.4	6.9	(+0.5)
10. Colleen	9.2	6.7	-2.5**
11. Curtis	9.2	9.2	(0.0)
12. Dale	9.6	7.7	-1.9**
13. Dean	6.9	5.5	-1.4*
14. Diana	8.5	9.1	(+0.6)
15. Donna	10.3	10.5	(+0.2)
16. Dorothy	10.5	10.6	(+0.1)
17. Douglas	7.7	9.0	+1.3*
18. Edgar	7.2	7.2	(0.0)
19. Elaine	11.1	8.8	-2.3**
20. Eugene	7.6	8.0	(+0.4)
21. Floyd	5.8	6.6	+0.8
22. Freda	9.7	10.7	+1.0
23. Gladys	8.0	6.7	-1.3*
24. Grace	6.4	6.1	(-0.3)
25. Guy	6.9	7.9	+1.0
26. Harvey	10.4	10.3	(-0.1)
27. Hazel	9.6	6.6	-3.0**
28. Helen	6.9	8.0	+1.1
29. James	7.5	7.5	(0.0)
30. John	10.6	10.6	(0.0)
31. Janet	10.8	10.7	(-0.1)
32. Joyce	8.4	9.7	+1.3*
33. June	5.9	6.9	+1.0
34. Leroy	7.1	7.0	(-0.1)
35. Louis	8.8	10.0	+1.2

STUDENT	ANTICIPATED ACHIEVEMENT GRADE PLACEMENT ^b	TOTAL READING GRADE PLACEMENT	DIFFERENCE ^c
36. Mabel	8.9	6.9	-2.0**
37. Mary	7.3	9.5	+2.2**
38. Nancy	9.3	8.9	(-0.4)
39. Phillip	9.7	8.3	-1.4*
40. Ralph	10.1	8.4	-1.7**
41. Robert	10.1	10.3	(+0.2)
42. Russell	9.7	8.8	-0.9
43. Sarah	10.9	10.7	(-0.2)
44. Susan	8.5	7.5	-1.0
45. Thelma	8.3	7.4	-1.1
46. Walter	11.4	11.0	(-0.4)
47. Warren	8.3	6.1	-2.2**
48. Wendell	8.3	7.4	-1.1
49. Wilma	10.9	9.6	-1.3*
50. Winifred	8.8	5.0	-3.8**
Total	436.8	416.1	
Mean	8.7	8.3	

^a As measured by the *California Reading Tests, Junior High Level*, administered during the first month of the school year (Monterey, Calif.: California Test Bureau, 1957).

^b Obtained from special tables for the Anticipated Achievement Grade Placement. (These are individualized expectancy norms for the *California Achievement Tests*.) For example, Albert's AAGP of 10.1 indicates that this is the average grade placement in reading comprehension for students in the norming sample in the same grade, with the same age and the same mental age as Albert.

^c Differences smaller than the $SE_{\text{difference}}$ (.78) have been enclosed in parentheses. One asterisk indicates a difference large enough to be significant at the 10 percent level, that is, a difference large enough to occur in only 10 percent of the sample testings if the true difference were zero. In other words, the odds are only 10 percent, or one out of ten, that the difference is due to chance and would not be found on retesting. Two asterisks indicate a difference large enough to be significant at the 5 percent level, with an even lower probability of representing a chance difference.

In constructing the scatter diagram in Figure 14.2, each student's identification or code number was entered in the square that located him with respect to both his AAGP (Anticipated Achievement Grade Placement) and his RCGP (Reading Comprehension Grade Placement). For example, student 13 (Dean) had an AAGP of 6.9 and an RCGP of 5.5; hence his code number (13) was entered in the square for an AAGP of 6.5-6.9 and an RCGP of 5.5-5.9 (that is, in the square formed by the intersection of the row and column in which pupils of Dean's ability level are recorded).

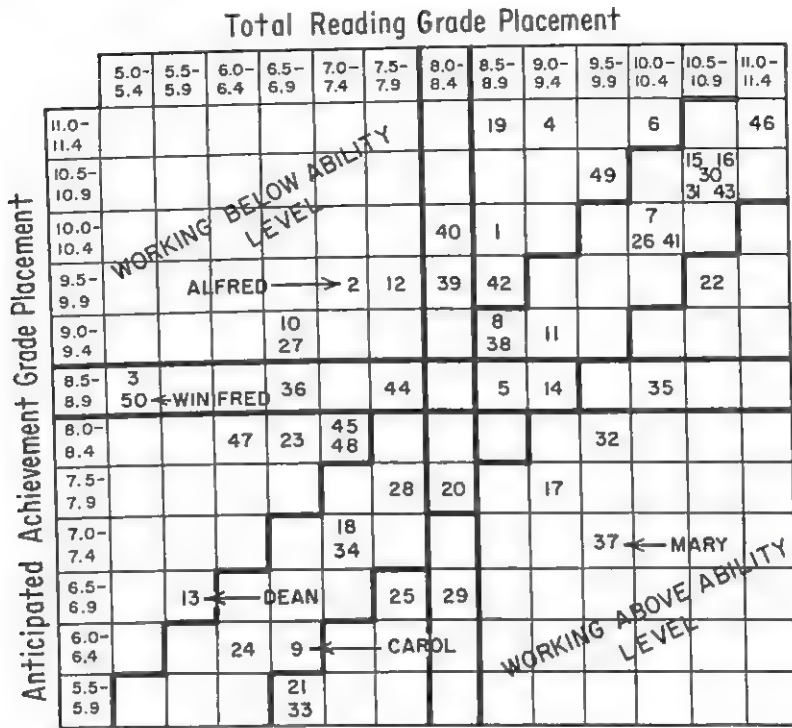


Fig. 14.2 Scatter-Diagram of Anticipated Achievement Grade Placements and Total Reading Grade Placements for 50 Ninth-grade students on the *California Reading Test: Intermediate* (Monterey, Calif.: California Test Bureau, 1957).

To aid in the interpretation of results, a pair of heavy horizontal lines has been drawn to enclose the class interval containing the mean (8.7) of the AAGP's. All students whose code numbers appear above this line are above average in expected achievement. The pair of heavy vertical lines enclose the class interval containing the mean reading comprehension grade placement of 8.3. All students with reading comprehension grade placements to the right of this line have achieved above the group average, whereas those to the left are below the group average.⁵ All students whose code numbers appear in the intervals to the left of 9.0 are below the national norm. Students in the upper left hand quarter of the scatter diagram are above average with respect to expectancy but below average with respect to achievement.

The "staircase" diagonal includes the code numbers of students who are

⁵ Since all data are grouped by intervals, students whose scores are within the interval in which the mean lies are considered to be *at* the average level, rather than either above or below.

working "approximately at expectancy level." It will be noted that in each case this channel includes the square for students whose AAGP and RCGP correspond (fall in the same interval) plus the adjoining intervals, representing a deviation of one half year in either direction. In this group of 50 pupils, 20 are in this center channel and can therefore be described as reading "approximately at expectancy level"; 9 are to the right of the channel, indicating that their reading is "above expectancy"; and 21 are to the left of the channel, or reading "below expectancy."

Among those who appear to be underachievers (those reading *below* expectancy), Alfred and Winifred have been chosen as examples. With AAGP's of 9.8 and 8.8, respectively, Alfred and Winifred have reading grade placements of 7.3 and 5.0, respectively. Obviously, these two students, in addition to many others in this group, need diagnosis and probably need corrective instruction in reading. Those students who are reading significantly below both the national norm and their own ability level can also be easily identified from this scatter diagram. These students can profit from remedial instruction much more than such students as Carol, who, although she reads well below national norm, is achieving up to expectancy.⁶ In interpreting an expectancy chart, the teacher should bear in mind that the data on both ability and achievement reflect errors in measurement.

USING EXPECTANCY CHARTS IN HIGH SCHOOL SUBJECTS By means of specially designed expectancy charts,⁷ it is possible to compare students' achievement in various high school subjects with their ability level. In Figure 14.3, the code number for each student is plotted in the column corresponding to his deviation IQ on the Otis, Pintner, or Terman-McNemar tests, and at a level in that column that corresponds to his achievement on the algebra test. For example, student 12 had an IQ of 88 on the Terman-McNemar test and a standard score of 117 in algebra (which is equivalent to a percentile rank of 80, as shown in the column to the right). When his code number is entered on the expectancy chart, it is seen that he would rank at the 98th percentile in comparison with students of his own ability level. In fact, as with Mary in the ninth-grade (Fig. 14.2), it seems advisable to obtain additional data to determine whether the intelligence quotient of this student has been underestimated by his performance on the Terman-McNemar test, used in plotting this chart.

⁶ In making this analysis, it is assumed that the intelligence test used in obtaining the AAGP has provided a valid measure of the student's mental age. If the student's reading handicap is sufficient to have invalidated his score on the test, administration of a nonlanguage intelligence test may be advisable.

⁷ Expectancy charts are available for tests of the *Evaluation and Adjustment Series* (a series of survey tests in the major high school subjects published by Harcourt, Brace & World, Inc.).

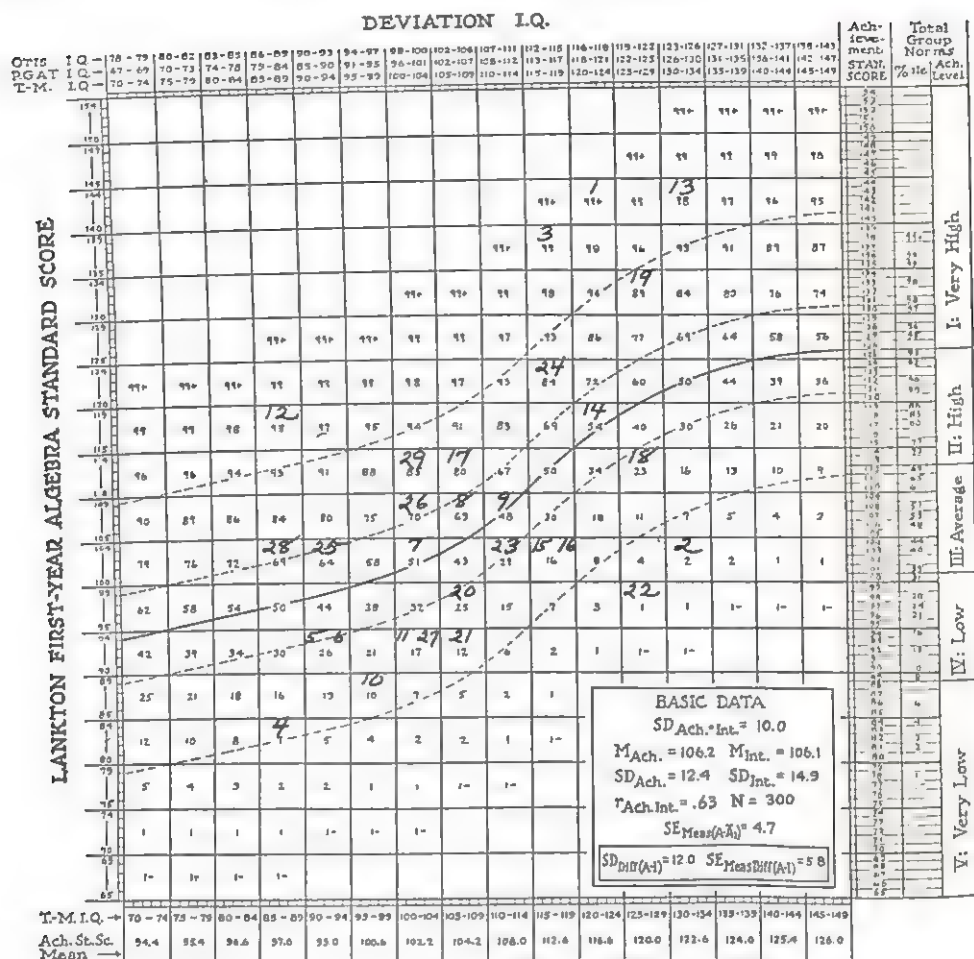


Fig. 14.3 Test Results for a Class Completing First-year Algebra, Charted on the Lankton First-Year Algebra Expectancy Chart. Achievement test results: standard scores on the Lankton test; intelligence test results: deviation IQ's on the Terman-McNemar Group Test of Mental Ability.

From *Manual, Lankton First-Year Algebra Test*. Copyright 1951 by Harcourt, Brace & World, Inc. New York, N.Y. Copyright in Great Britain. All rights reserved. Reproduced by permission.

Students 1, 3, and 13 are all high-ability students whose achievement in algebra is even better than would be expected from persons of their learning ability. They merit special encouragement to continue in mathematics, and the algebra teacher may wish to plan enrichment activities for them. Students 2 and 22 also have superior ability but are achieving far below their potential. A little time spent in diagnostic study of the reasons for their low achievement would probably bring good results, since these stu-

dents should be able to make excellent progress with a minimum of teacher assistance. Those students with IQ's above 100 and achievement below average for their ability level could profitably be assigned diagnostic testing material and given special help as a group with their common problems or learning difficulties. The advisability of continued instruction in algebra for students 4, 5, 6, and 10 might well be discussed with the high school counselor, who would have additional data on learning ability, motivation, and career plans.

Locating the Errors or Learning Difficulties

If a student has been identified on a standardized test as working below expectancy in some area, for example, arithmetic fundamentals, the teacher's next step depends upon the type of standardized test he has used. A few standardized tests (for example, the *California Achievement Tests*) provide subtest scores in addition, subtraction, multiplication, and division, which help the teacher to determine specific areas of difficulty for the student. By means of the Diagnostic Analysis printed in the test booklet, the teacher can obtain still further leads concerning the types of error made. If the Scoreze device has been used, he can easily identify the specific problems missed and, by referring to the printed descriptions, can classify them by type.

Study of the Diagnostic Analysis and the Scoreze⁸ for an individual student might suggest that he had mastered the combinations but had difficulty with problems involving zeros, or that he consistently failed in problems requiring inversion of the divisor in the division of fractions. Such an analysis, however, provides only clues or leads. The teacher must follow up such leads by assigning similar problems and noting the student's performance on a larger sampling of problems of the same type.

For students who are markedly retarded in the fundamental skills, a more comprehensive diagnosis of learning difficulties is desirable as a basis for a well-planned program of corrective instruction. For this purpose, special diagnostic tests are available, designed to give evidence on students' specific retraining needs.⁹

⁸ For an illustration of the use of the Diagnostic Analysis and Scoreze for an individual pupil, see Theodore L. Torgerson and Georgia Sachs Adams, *Measurement and Evaluation for the Elementary School Teacher* (New York: Holt, Rinehart and Winston, Inc., 1954), pp. 206-208.

⁹ It is only in the fields of reading and arithmetic, however, that much work has been done in the development of published diagnostic tests. A number of specific tests and techniques that can be used for diagnosis in the skills subjects are presented in Chapters 11 through 14 of the authors' textbook for elementary school teachers and Chapters 13 through 16 of the authors' textbook for secondary school

The subtests of a diagnostic test should measure component skills that are critical for success in the subject, for example, word attack skills in reading. Each subtest on a component skill should be loaded with opportunities for one type of error and be as free as possible from other sources of learning difficulty. Ideally, each subtest should be long enough so that intraindividual differences between related abilities can be reliably measured. The student will recall that when tests are highly correlated, the differences between scores tend to have low reliability.

Since a high level of reliability is seldom achieved in diagnostic tests, the user should interpret the results cautiously, making tentative hypotheses about learning difficulties and appropriate corrective instruction. The teacher must be willing to revise such hypotheses as additional information is obtained and must search for additional clues if corrective instruction proves ineffective.

Ideally, diagnostic tests should be directly related to the materials available for corrective instruction. They should always be gauged to the achievement level of the retarded student, that is, with a large number of items at his achievement level. Norms are relatively unimportant, since the chief purpose in diagnostic testing is to discover what the student cannot do and why, rather than to compare his achievement with grade standards.

The teacher who administers diagnostic tests to students with learning problems not only obtains data that aid these students being studied but increases his own understanding of the learning process and of typical errors and confusions among retarded students. In other words, the teacher who uses and interprets diagnostic tests obtains valuable in-service training, which may not only improve his teaching and his selection of practice materials but may help him to interpret his everyday observations of students during supervised study periods and may change his correction of homework from a routine clerical task to one that has real diagnostic value.

DIAGNOSING LEARNING DIFFICULTIES IN ARITHMETIC Ideally, diagnostic tests in arithmetic are keyed to general screening tests, as well as to individualized practice materials. For example, a teacher using the Brueckner series¹⁰ can administer one or more of the screening tests (available for whole numbers, fractions, and decimals). Then, on the basis of results on these screening tests, he can select appropriate diagnostic tests from the

teachers. T. L. Torgerson and Georgia Sachs Adams, *Measurement and Evaluation for Elementary School Teachers* (New York: Holt, Rinehart and Winston, Inc., 1954); Georgia Sachs Adams and T. L. Torgerson, *Measurement and Evaluation for Secondary School Teachers* (New York: Holt, Rinehart and Winston, Inc., 1956).

¹⁰ Lee J. Brueckner, *Diagnostic Tests and Self-Helps in Arithmetic* (Monterey, Calif.: California Test Bureau, 1955).

23 such tests in the series. On the basis of the results from such individualized testing, students can be assigned to work on any of 23 sets of corrective "self-help" exercises, keyed to the diagnostic tests; and other suggested remedial procedures can be used.

For some children, printed tests and self-help materials are not adequate; for example, a student may not have the reading ability and/or the motivation level to work alone on such materials. For such children, it may be necessary to have the student work the problems aloud. In this way, one can study the process as well as the result. In observing students during supervised study, the teacher may note evidence of counting or other roundabout methods of computation.

A few moments of observation at a critical time in the introduction of a new process may forestall later difficulties. Through studying the work of retarded students on their regular assignments, the teacher may be able to spot recurring errors and misconceptions.

Problem solving in arithmetic requires competence not only in *computational skills* but in many other skills as well. A command of reading skills is essential in problem solving. Treacy found that students who were poor in problem solving tended to be deficient in both general vocabulary and arithmetic vocabulary, as well as in four subtests of the *Diagnostic Examination in Reading Abilities*.¹¹

In diagnosing a student's difficulties in problem solving, the teacher will wish to note:

1. his test scores in work-type silent reading;
2. his knowledge of arithmetic vocabulary and symbolism (as revealed in subtests of the *SRA Achievement Series*, the *California Achievement Tests* or other standardized tests;
3. his scores on exercises in problem solving (selecting the process to be used, estimating answers, and the like), such as those provided in the above-mentioned tests.

Before undertaking a thorough diagnosis, however, the teacher should ascertain the extent to which errors in arithmetic problems may be due to errors in computation, carelessness in arranging work, and general lack of neatness. For these causal factors, drill in arithmetic combinations plus improved motivation may be all that is needed. A more thorough diagnosis is not required unless the student shows evidence of failure to grasp the quantitative relations involved or generally inefficient methods of problem solving.

¹¹ John P. Treacy, "The Relationship of Reading Skills to the Ability to Solve Arithmetic Problems," *Journal of Educational Research*, vol. 38 (October 1944), pp. 86-96.

DIAGNOSING LEARNING DIFFICULTIES IN READING After the students who need special help in reading have been identified, diagnostic tests should be administered to determine the specific nature of students' difficulties. Results on the Survey section of the *Diagnostic Reading Tests*, for example, will help the teacher to decide which of the diagnostic tests should be given. For students whose general level of reading comprehension is below their ability level, both parts (Silent and Auditory) of Section 2 (Comprehension) should be administered. For students who attain low reading-rate scores on the Survey test, more specific diagnostic information can be obtained by the administration of Section 3 (Rates of Reading), which provide data on the student's reading rate in two subject-matter areas, as well as his ability to read rapidly under pressure. Finally, for students who have scored low in vocabulary and silent reading as compared with auditory comprehension, the teacher may wish to administer individually Section 4, on Word Attack. In the first part of this test, the student is asked to read aloud six graded paragraphs of interesting general-type reading material. The teacher observes the student's reading attitude and methods, and records all errors, using the recommended notations for substitutions, omissions, repetitions of two or more words, mispronunciations, insertions, and the like.

The teacher of a remedial-reading class may need to know the words in a primary sight vocabulary that a student has not learned or the specific skills in word attack that he has not mastered. Under such circumstances, *individual* administration of a word-perception or oral-reading test may be indicated. A student's performance in oral reading reveals his mastery of the basic skills of word recognition and word analysis. When a pupil reads orally, the teacher can note his reading habits and his fluency in reading, as well as the kinds of errors he makes in pronunciation.

An oral-reading test is usually administered by having a student read each paragraph aloud while the examiner evaluates his performance, recording such errors as hesitations, insertions, mispronunciations, omissions, repetitions, and substitutions. If a student's oral-reading performance is recorded on tape, more objective analysis and scoring are possible. In some oral-reading tests, the student's comprehension of the material read is checked by having him respond orally to standard questions read to him by the examiner. On some of the tests, the student's rate of reading is also noted. Three illustrative oral-reading tests will be briefly described.

1. The *Gates Diagnostic Reading Tests*¹² are individually administered oral-reading tests. Two forms are available. The series consists of a pupil's record booklet for the teacher, two sets of cards containing test material for the

¹² Arthur I. Gates, *Gates Reading Diagnostic Tests* (New York: Bureau of Publications, Teachers College, Columbia University, 1945).

child, and a manual that contains directions, grade and age norms, and suggestions for remedial procedures. From the 18 tests one can choose those likely to help in analyzing the child's specific difficulties.

2. The *Durrell Analysis of Reading Difficulty*¹³ for grades 1 to 6 consists of a set of eight graded paragraphs to measure oral reading habits, another set of eight paragraphs for measuring oral recall, and two additional sets of paragraphs of comparable difficulty to be read silently in order to measure oral and written recall. The series also contains a total of 175 words to be flashed in a cardboard tachistoscope, in order to measure word recognition and word analysis. An accompanying record booklet contains several classified checklists of reading difficulties.
3. The *Gilmore Oral Reading Test*¹⁴ consists of ten oral-reading paragraphs that form a continuous story. The paragraphs are scaled in difficulty, and each is accompanied by five questions to check comprehension. The test is designed to measure accuracy of oral reading, rate of oral reading, and comprehension in grades 1 through 8.

For suggestions for diagnostic work in handwriting, language usage, foreign language, and several other subject areas, the student is referred to the authors' textbooks for elementary and secondary school teachers.

Discovering the Causal Factors

In some cases of learning difficulty, the causal factors are relatively simple. A student may (as the result of inattention, insufficient or inefficient practice, or irregular attendance) have failed to learn basic vocabulary or verb forms in foreign language or to understand the process of multiplying signed numbers in algebra. If such causal factors are temporary, multiplying signed numbers in algebra. If such causal factors are temporary, it is sufficient to identify and remedy the gaps in the student's learnings. If, however, inattention or any other behavior is a *persistent* factor, it may be symptomatic of underlying difficulties. In these situations, attempts must be made to identify and remedy the basic causes. Until this is done, corrective instruction will remain at the "patching up" level of effectiveness.

If a student consistently achieves below expectancy in most of his subjects, certain basic causal factors, such as poor health, faulty work habits, or emotional maladjustment may be involved. Although no single factor may seem to be serious, the combined effect of several factors may produce significant scholarship and behavior problems.

Although the basic causes of low achievement for a given student are usually complex and interrelated, most of them can probably be classified under five major categories:

¹³ Donald D. Durrell, *Durrell Analysis of Reading Difficulty* (New York: Harcourt, Brace & World, Inc., 1937).

¹⁴ John V. Gilmore, *Gilmore Oral Reading Test* (New York: Harcourt, Brace & World, Inc., 1952).

DISABILITIES IN THE BASIC SKILLS Retardation in many subjects can frequently be attributed to retardation in the basic skills of reading and arithmetic. Difficulties in social studies, science, and many other subjects may be due to an inability to read textbook materials with comprehension. Difficulties in arithmetic reasoning or in the higher processes of arithmetic computation are attributable, in part, to the fact that the basic arithmetic skills are not functioning at an automatic level. Difficulties in algebra, chemistry and physics, industrial arts, and other subjects may be due, in part, to deficiencies in arithmetic.

Corrective instruction in the basic skills is necessary if the students are to get adequate returns from their study time in the subjects in which these skills are constantly demanded.

INADEQUATE WORK-STUDY SKILLS Many students fail to do their best work because of inadequate methods of attacking learning problems. Many subjects place a premium on the student's ability to find information in books, to locate related source materials, and to read maps, graphs, and tables. The student must know the technical vocabulary of a subject field if he is to comprehend the content.

SCHOLASTIC APTITUDE FACTORS Although some authors list "inadequate mental maturity" as a major cause of learning difficulties, it is obviously the *disparity* between the abilities required in the teaching-learning situations and the student's mental maturity that is at fault. In other words, retardation and discouragement may result from the student's having been programmed into subjects that are too difficult for him or from the use of instructional materials that are too difficult.

The modern school has accepted the responsibility of helping each student achieve to the level of his capacity. The limitations of group instruction, however, make it desirable that the student's mental maturity not deviate too markedly from the average of his class or his instructional group within the class. When learning tasks are too difficult, frustration destroys interest and incentive; when they are too easy, boredom can lead to minimum involvement and inadequate effort.

PHYSICAL FACTORS Chronic diseases, impairment of vision and hearing, and other physical handicaps interfere with learning. Lowered vitality, distractibility, and irregular attendance hinder learning and decrease retention. The student's resulting discouragement and lack of interest may contribute, in turn, to further retardation.

The cumulative health records of inefficient learners should be checked and interpreted with the aid of the school nurse. Such a record not only

indicates a student's health status as of the time of his latest medical examination but also provides a longitudinal picture of his physical development. Significant physical characteristics are recorded there as well as recommendations that have been made to parents and educators concerning necessary remedial or preventive measures. In schools that use the Wetzel "grid" (a graphic record form for cumulated height-weight data), significant "growth failures" (or deviations from a student's expected growth pattern) can be readily identified. In a study of more than 2000 students, the Wetzel grid identified 95 percent of students classified as "poor" or "borderline" by school physicians.¹⁵

Impaired vision is a serious hazard to learning. The Snellen Chart, the most widely used screening test, fails to detect moderate farsightedness or astigmatism, as well as difficulties in eye coordination; in fact in one study, use of the Snellen Chart detected eye defects in only about half of the students who showed defects in a thorough ophthalmological examination.¹⁶ Hence, many school systems use a telebinocular to measure eye-muscle balance, depth perception, visual fusion, and other abilities important to effective reading; even these tests, however, are intended only to identify students who should be referred to ophthalmologists for further study.

Defective hearing is a serious hazard to learning. The widespread use of audiometers for testing students' hearing has greatly increased our effectiveness in early identification of hearing loss. As many as 40 students can be tested at a time; students who do poorly on a group audiometer test can be tested individually with an instrument that checks acuity of hearing at different pitches.

EMOTIONAL FACTORS The emotional tensions of the poorly adjusted student may affect his concentration, motivation, and persistence of effort. A student's fear of failure may almost paralyze his efforts; his self-consciousness may cause him to withdraw from participation in class activities; as a result of his hostility toward adult authority, he may refuse to do assigned tasks; his anxiety to do things well may prevent him from developing reasonable speed in reading, handwriting, and other skills; continued frustrations may result in a retreat to daydreaming and psychological deafness to the teacher's voice. At adolescence, preoccupation with social status and boy-girl relationships may cause a drop in achievement. Methods of studying emotional factors have been considered in Chapters 8 and 9.

¹⁵ Norman C. Wetzel, "The Simultaneous Screening and Assessment of School Children," *Health and Physical Education*, vol. 10 (December 1942), pp. 576-577.

¹⁶ T. H. Eames, "The Effect of Correction of Refractive Errors on the Distant and Near Vision of School Children," *Journal of Educational Research*, vol. 36 (December 1942), pp. 272-279.

GROUP DIAGNOSIS

The use of standardized tests in group diagnosis was discussed in Chapter 13. Almost any measurement procedure used by teachers can be used in diagnosis. For example, an algebra teacher introduced the principles involved in the addition, subtraction, multiplication, and division of signed numbers, devoted a week to teaching them, and then gave a ten-item test covering what had been taught. Although this test was much too short for individual diagnosis, the results could aid in group diagnosis. By use of the show-of-hands method (discussed in Chapter 10), the teacher tabulated the errors for each item, as follows:

ITEM	ERRORS
1	2
2	1
3	3
4	2
5	3
6	10
7	8
8	9
9	32
10	32

From this simple tabulation, the teacher was able to see that most of the difficulty lay in the last five examples. Since examples 9 and 10 had been missed by everyone, it was obvious that the concept they involved—subtraction of one negative number from another—needed to be retaught to the entire class. The errors on examples 6, 7, and 8 involved the multiplication or the division of two negative numbers. The teacher could ask students who missed these problems to come to the chalkboard during a supervised study period; he could then give further explanation, answer their questions, and require them to work additional problems of this type.

A simple group diagnostic analysis of this sort allows the teacher to locate quickly the specific items on which many of the students are having trouble. These items can then be studied to identify what is causing trouble. Review and reteaching can then be focused on the learning needs of the students.

If a test is designed for machine scoring and an item-count attachment is available on the local IBM test-scoring equipment, a report such as the one shown in Figure 14.4 can be easily obtained. This chart summarizes

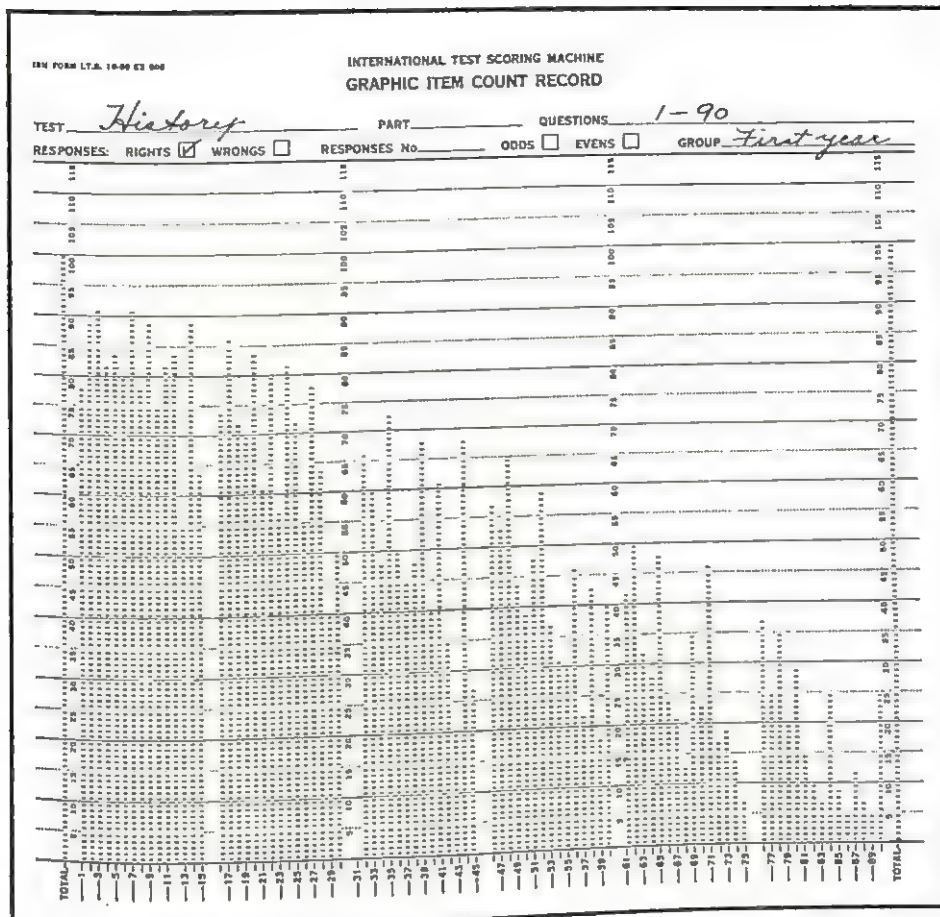


Fig. 14.4 Item Count of the Number of Correct Responses of 100 Students to Each of 90 Items of a History Test.

Reproduced by permission of the International Business Machines Corporation.

student successes on items 1 to 90 of a departmental history examination. Prepared by counting student responses electronically, an analysis can be made of a hundred papers in ten minutes. If the test has more than 90 questions, the process must be repeated for each group of 90 or fewer items.¹⁷

¹⁷ If the teacher wishes to have separate analyses for the high-scoring or low-scoring halves of his group (or the high-scoring and low-scoring fourths), he need only group the papers and request that the scoring machine operator record results for each group on a separate graphic item count record. Then, item analysis results are easily transferred to the test itself, or to test-item cards in the manner suggested in Chapter 10.

It is evident from Figure 14.4 that several questions were so easy that they had little measurement value. Others were missed so frequently that they may indicate a need for reteaching the concepts involved. (The possibility that a question is difficult because of ambiguity should always be considered.) If a teacher is using a standardized or departmental test, he may wish to go through the test *before* the item count is made and mark the questions that he feels have curricular validity for his instructional program. Then, from the results of the item analysis, he can see how well his students succeeded on those test questions that he considers to be valid and significant.

In the manuals for many standardized achievement tests,¹⁸ data are given regarding the percentage of students in the representative norming sample who answered each item correctly. The teacher can compare the results of his own item analysis with these data in order to determine for each item the relative standing of his class or classes with respect to these percentage-correct norms. This comparison has added significance if the teacher has first evaluated all items of the standardized test in terms of their curricular validity for his own instructional program.

BASIC PRINCIPLES OF CORRECTIVE INSTRUCTION

Three essential steps in educational diagnosis have been explored: (1) identifying the students who are having trouble, (2) locating the errors or learning difficulties, and (3) discovering the causal factors. The results of such diagnosis have significance only if they constitute the basis for corrective instruction and for remedial procedures that remove, alleviate, or compensate for causal factors in the student and his environment.

It is not within the scope of this book to outline corrective procedures for each of a wide variety of learning difficulties and causal factors. In fact, the learning process and the dynamics of human motivation are so complex that it is undesirable, if not impossible, to match corrective procedures to learning problems in any rote or mechanical fashion.¹⁹

¹⁸ For example, the tests of the *Evaluation and Adjustment Series* (Harcourt, Brace & World, Inc.), and the STEP tests (Educational Testing Service).

¹⁹ General suggestions, and selected lists of remedial materials in several subject fields are given in Chapters 11 through 14 of the authors' textbook for elementary school teachers and Chapters 13 through 16 of the authors' textbook for secondary school teachers. T. L. Torgerson and Georgia Sachs Adams, *Measurement and Evaluation for Elementary School Teachers* (New York: Holt, Rinehart and Winston, Inc., 1954); Georgia Sachs Adams and T. L. Torgerson, *Measurement and Evaluation for Secondary School Teachers* (New York: Holt, Rinehart and Winston, Inc., 1956).

An important distinction should be made between corrective instruction in *information and understandings* and corrective instruction in the *basic skills*. If a teacher can identify several students who lack a thorough understanding of certain concepts (for example, latitude and longitude), he may reteach these concepts through group instruction, demonstrations, and supplementary reading by the students. *General* retardation in the content subjects, however, is frequently due to inadequate mastery of the basic skills of reading, arithmetic, or language, or to inadequate command of the work-study skills. Hence, corrective work in the basic skills plus improved motivation in the content subjects may be sufficient to effect improvement.

Deficiencies in the basic skills of reading, arithmetic, and language can be corrected only in part by special group instruction or by individualized assistance during supervised study. In these fields, the best results can be achieved only by systematic, meaningful practice on instructional materials designed to develop the specific skills that the individual has failed to master.

Selection of Materials

Selection of corrective materials for a student is a crucial aspect of his corrective instruction. Any materials selected should meet the following criteria:

1. The difficulty of the corrective materials should be geared to the student's readiness or maturity in the subject or skill to be improved. If the student's interest is to be maintained, corrective instruction must result in feelings of accomplishment. The student's grade equivalent on a standardized test may be used as a partial basis for selecting the level of instructional materials to be used. A set of remedial materials without grade labels, which provides for a wide range of difficulty, should be used.
2. The corrective materials should be designed to correct the student's individual difficulties. By means of observation, interview, and diagnostic testing materials, the teacher will have analyzed the work of the retarded student in order to locate his specific retraining needs. An adequate amount of corrective material, designed to correct the specific difficulties discovered, should be provided.
3. The corrective materials should be largely self-directive. An individualized instructional program cannot achieve optimum effectiveness unless the materials are self-directive, permitting a number of students to work independently on different materials. Written directions, easily read and understood by the students, must accompany the materials, so that a minimum of direction by the teacher is required.
4. The corrective materials must permit *individual rates of progress*.
5. A method should be provided for recording evidence of individual progress. When the student has an opportunity to record his successes on a progress record, he is given an additional incentive to achieve.

The reader will recognize that well-designed programmed instruction meets these criteria.

Planning and Carrying Out the Program

Although the selection of remedial materials is highly important, it is only one aspect of the teacher's attack upon learning difficulties and underlying causative factors. The following principles should guide the teacher in planning and carrying out the program:

1. One of the first steps should be the correction of any physical factors that affect learning, such as defects of hearing or vision.
2. The cooperation of the parents should be obtained in correcting such physical factors, alleviating emotional tensions, providing better study conditions, and the like.
3. If the student seems to have little desire to learn, immediate steps should be taken to try to improve his attitude through providing activities in which he can enjoy success, receive praise for his efforts, and be given opportunities to develop his special interests or use his special skills. Personal interviews may do much to establish rapport and provide leads that will help the teacher know each student's interests and problems.
4. Corrective instruction should begin by analyzing *with the student* his specific strengths and needs and showing how the instructional materials are designed to correct *his* deficiencies. When the student is helped to face his problems constructively and provided with aids to solving them, he can usually take the first steps that lead to early evidence of progress.
5. Instruction should begin at, or slightly below, the learner's present level of achievement. Short-term goals should be established which the learner considers reasonable and possible of attainment. By means of progress charts, praise, and social recognition, the student's feeling of successful accomplishment should be reinforced.
6. Since corrective instruction must usually proceed on the basis of a tentative diagnosis, the teacher must be ready to modify the remedial program if the approach and materials selected seem to be ineffective.
7. The results of corrective instruction should be evaluated—that is, comparable forms of a standardized test should be administered before and after a period of concentrated instruction. The effectiveness of the program must be evaluated *for each student* rather than merely in terms of class averages.
8. A record should be made of the results of each student's diagnosis, of methods and materials used, and of the results of corrective instruction. Such a record is not only helpful in the determination of next steps; it is likely to be invaluable to the next teacher if the student continues corrective instruction.

In individualized instruction, the teacher is constantly reminded of a principle that he frequently overlooks in other teaching situations—that is, that *learning* rather than teaching is the goal of his activities. As emphasized earlier, the growth of each individual—rather than the change in

group averages—is the criterion of success. Hence, the teacher needs a rich background in psychology and educational diagnosis, as well as consultant help from specialists, in order to attack successfully the variety of individual problems that present themselves.

SUMMARY STATEMENT

The first step in diagnosis is to identify the students who require further study. Low achievement may result chiefly from limited learning ability. It is therefore desirable to compare a student's achievement with some measure of expected achievement or with the distribution of test scores for students of his mental maturity level, as shown on an expectancy chart. Low achievement calls for detailed diagnosis only if it is significantly below expectancy for the student.

The second step is to locate and study the specific errors or difficulties, using tests that are valid and reliable for this purpose. Testing can be followed by a teacher-diagnostic interview to discover how and why the errors were made. Teacher-made tests can also be used as diagnostic instruments, provided the teacher constructs tests that provide many opportunities for students to make crucial errors and takes the time to analyze the specific mistakes made by the students.

The third step is to discover the causal factors. Causes of poor learning can be grouped under five general headings; disabilities in the basic skills, inadequate work-study skills, scholastic aptitude factors, physical factors, and emotional factors.

Corrective instruction in the skills should be individualized, permitting each student to work independently at his own rate on materials that have been selected or designed to correct his specific deficiencies. Appropriate materials for individualized instruction should provide for self-direction and be of the proper difficulty to ensure successful performance. Motivation should be improved through providing success experiences, praising efforts, and developing good rapport. The results of corrective instruction should be evaluated in terms of growth for each student, and a record should be made of methods and materials used and gains effected.

SELECTED REFERENCES

- BLIESMER, EMERY P., "Methods of Evaluating Progress of Retarded Readers in Remedial Reading Programs." *15th Yearbook, National Council on Measurements Used in Education*. New York: The Council, 1958, pp. 128-134.
- BRUECKNER, LEO L., "Diagnosis in Teaching," in C. W. Harris, ed., *Encyclopedia of Educational Research*. New York: The Macmillan Company, 1960.
- _____, AND GUY L. BOND, *The Diagnosis and Treatment of Learning Difficulties*. New York: Appleton-Century-Crofts, 1955.
- KIRK, BARBARA A., "Test Versus Academic Performance in Malfunctioning Students," *Journal of Consulting Psychology*, vol. 16 (June 1952), pp. 213-216.

- KOUGH, JACK, AND ROBERT F. DEHAAN, *Helping Children With Special Needs. Teacher's Guidance Handbook, Part II, Elementary Edition*. Chicago: Science Research Associates, Inc., 1956.
- , AND ———, *Identifying Children Who Need Help. Teacher's Guidance Handbook, Part I, Elementary Edition*. Chicago: Science Research Associates, Inc., 1955.
- LINDQUIST, E. F., ed., *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, Chapter 2.
- TRAXLER, ARTHUR E., "The Use of Tests in Differentiated Instruction," *Education*, vol. 74 (January 1954), pp. 272-278.
- TRIGGS, FRANCES O., AND OTHERS, *Diagnostic Reading Tests: Their Interpretation and Use in the Teaching of Reading*. New York: Committee on Diagnostic Reading Tests, Inc., 1952.
- WOOD, ERNEST R., "Subject Disabilities: Special Difficulties in School Learning," in Charles E. Skinner, ed., *Educational Psychology*, 3d ed. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1951, pp. 484-521.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. Consider the tenets discussed in this chapter, and identify those that were followed in the schools you attended as a pupil.
2. Compare survey and diagnostic tests with respect to the purposes they serve. Cite a specific test of each type in one subject field.
3. What additional data would you like to have concerning Alfred and Winifred (underachievers in Fig. 14.2) in order to develop hypotheses as to why they are not working up to ability level?
4. Analyze and evaluate a diagnostic test in silent reading (or arithmetic) and indicate whether differences between subtest scores can be interpreted with confidence.
5. What aspects of reading disability may be revealed through the use of an oral reading test?
6. Select a diagnostic test in arithmetic and describe the skills and abilities that it is purported to measure.
7. Evaluate several specimens of handwriting on the basis of a diagnostic handwriting scale.

PART FOUR

Administrative,
Supervisory,
and Guidance Aspects

Planning and Administering the Evaluation Program

In Part Three, major emphasis was placed on the use of tests and other evaluation techniques by the classroom teacher. Since teachers have considerable freedom in the planning of educational experiences for their students, they must inevitably assume greater responsibility for evaluating the worthwhileness of those experiences. Moreover, to the extent that schools are committed to the ideal of meeting student needs and individualizing instruction in terms of those needs, the teacher becomes the key person in interpreting measurement data and in putting such data to use in individualizing instruction and in guiding students.

Certain aspects of the modern evaluation program, however, are school-wide and system-wide in scope. For example, the objectives that the teacher develops as a guide to his own evaluation activities should be in harmony with the common objectives of the educational program, as cooperatively developed on a school-wide or system-wide basis. Decisions must be made on at least a school-wide basis concerning the types of measurement data that should be obtained and recorded for *all* students. Agreement should be reached, also, on the most valid, reliable, and practicable instruments to be used for obtaining such data. In the selection and development of diagnostic tests, as well as a variety of teacher-made tests, rating scales, and other evaluation instruments, it is desirable for groups of teachers to coordinate their efforts. In-service education is needed in the administration, scoring, and interpretation of tests, as well as in the use of the more informal evaluation techniques, in order that the results will be valid and reliable.

In this chapter we are concerned primarily with the *school* evaluation program as it involves (1) the cooperative development and clarification of school objectives; (2) the administration and use of standardized tests

in obtaining certain data for all students; and (3) the cooperative selection and development of supplementary evaluation materials to be available to all teachers as needed.

FUNCTIONS OF THE EVALUATION PROGRAM

Although the distinctions among the functions are by no means clear-cut, data provided by an evaluation program can be used to implement certain (1) administrative-supervisory, (2) instructional, and (3) guidance functions.

Administration and Supervision

From the *administrative-supervisory* point of view, measurement data for class and grade groups can assist in determining needs for supervisory assistance or for a revision of curricular activities and instructional materials. Measurement data are indispensable in evaluating student progress toward educational goals and reporting on such progress to the governing board and the community. The measurement program also affords data needed for the maintenance of the cumulative records essential in the transfer and promotion of students. Measurement data aid in the making of administrative decisions concerning the classification of students, enabling the administrator to reduce the heterogeneity of classes or to identify students to be assigned to special remedial groups. Tests are indispensable aids in the selection of students by college admissions officers.

Instruction

Evaluation serves four major *instructional* functions: (1) to determine the extent to which students are making progress toward instructional goals; (2) to provide evidence to students and their parents concerning such progress; (3) to ascertain group and individual retraining needs as a basis for grouping within the class, and individualized corrective instruction; and (4) to aid in developing hypotheses concerning basic causal factors behind any deficiencies. In carrying out the first and second functions, the teacher is concerned with measuring both short-term gains, as the result of specific units of instruction, and long-term gains over a period of a semester or year.

Guidance

The *guidance* functions of evaluation are closely interrelated with the instructional functions. They are chiefly concerned with the use of evalua-

tion data in (1) helping the counselor to assist students in making wise decisions in the areas of educational and vocational guidance; (2) helping counselors and teachers to understand the students' needs in the area of personal-social adjustment; (3) helping students toward improved self-appraisal and self-direction; (4) increasing the parent's understanding of the special strengths and needs of his son or daughter as relevant to needed decisions; and (5) assisting in the identification of students whose special abilities or disabilities require unusual modifications in the educational program or referral to specialists.

CHARACTERISTICS OF AN EFFECTIVE EVALUATION PROGRAM

The functions of evaluation listed above cannot be achieved by an uncoordinated, hit-or-miss program or by a testing program limited to the measurement of the basic skills. In order to serve all functions effectively, an evaluation program should meet the following criteria:

1. It should be *based on a realistic statement of educational objectives*. The educational program is designed to bring about certain changes in student behavior—to teach students to do certain things better, to better understand concepts and principles, to make certain choices more wisely, than if they had not had the educational experiences provided. The evaluation program should provide evidence on the extent to which these changes in student behavior have been achieved.
2. Evaluation should be *comprehensive*. Plans should be made for evaluating student growth toward each major objective. Important goals should not be omitted in evaluation just because progress toward them cannot be appraised with high objectivity. "The importance of the outcome, rather than the precision with which its attainment can be measured, governs the effort to be devoted to obtaining and appraising the evidence."¹
3. The analysis of educational objectives and the subsequent planning of the evaluation program should be a *cooperative* undertaking. All teachers should participate in the group thinking required to establish and analyze goals and in the selection or development of instruments used in evaluation. At the classroom level, students should participate in analyzing objectives as they see them, in setting up criteria for evaluating achievement in specific activities, and in appraising individual and group products. Parents should also participate in planning those aspects of the evaluation program that are concerned with the reporting of progress to the home.
4. Evaluation should be a *continuous* process, providing a longitudinal picture of each student's growth rather than an occasional cross-sectional survey of current status. That is, evaluation should be used in assessing through pre-

¹ Warren G. Findley, "Gain vs. Status Scores as Evidence of Effectiveness of Instruction," *19th Yearbook, National Council on Measurement in Education* (Ames, Iowa: The Council, 1962), p. 45.

tests students' previous learnings, in appraising student growth in short units of instruction, in revealing to students the specific inadequacies in their achievement, in accumulating for the teacher the information necessary to plan next steps in instruction, in providing an objective basis for reporting to students and parents, and in serving many other functions that are integral parts of the instructional process.

5. The evaluation program should be *flexible*. Certain tests can be used to advantage on a school-wide basis and at scheduled times. Still other evaluation instruments should be available to teachers to be used whenever they contribute to the instructional program. Many instruments, such as student self-rating scales, will be developed in the classroom by teachers and students working and thinking together. Diagnostic tests and remedial materials should be available for use with individual students as needed.
6. All evaluation activities should be *keyed closely to the local situation*. Although many published tests will be usable, local norms and expectancy tables should be developed. Some instruments will have to be developed locally. Unless the content of a published achievement test is closely related to the objectives emphasized in the local curriculum, it should not be used to measure outcomes just because it has prestige or is readily available.
7. The functions of the evaluation program *should not be limited to an appraisal of group progress*. Most of the tests used in an evaluation program should have sufficient validity and reliability to be useful in guiding individual students and appraising their progress.
8. The evaluation program should utilize a *variety of techniques*. Evaluation is not limited to the use of paper-and-pencil tests; it includes all means of obtaining valid data concerning desired changes in student behavior.

PLANNING THE EVALUATION PROGRAM

If school evaluation programs are to meet the criteria listed above, much cooperative effort is needed. In the typical school system, practice lags far behind accepted theory. Measurement, evaluation, diagnostic study, and corrective instruction should not be considered supplementary activities to be performed if teachers find time to do so. They are essential aspects of effective teaching, which facilitate learning and should become an integral part of the instructional activities of every teacher.

The planning of a functional evaluation program must be based upon cooperative decisions reached within the school on (1) what types of judgments and decisions need to be made *about* students and *by* students and (2) the general and specific objectives of the educational program. The answers to the first question will help in the selection of tests to be used in making institutional decisions and helping students with their educational and vocational choices. Answers to the second question will aid in planning an evaluation program that will help to assess student progress and evaluate the effectiveness of the instructional program.

The necessary steps in planning an evaluation of the instructional program include: (1) the cooperative development of a realistic list of objectives for the educational program; (2) the analysis of these objectives in terms of changes in student behavior; (3) exploration of possible techniques and instruments with which to obtain evidence concerning changes in student behavior; and (4) the selection and development of the actual instruments to be used.

Cooperative Development of Objectives

Although many excellent lists of objectives are available, they tend to be so generalized that they provide little guidance in the selection of curricular activities and the development of evaluation instruments. The principal of each school, with the aid of subject supervisors, should lead his teachers in the cooperative development of a realistic list of intermediate objectives in each subject area that are related to the ultimate objectives of the educational program.

The process of group thinking that a teaching staff goes through in stating realistic objectives for its own students is an exceedingly valuable one. Goals that have seemed to be only attractive slogans take on real meaning, and their implications for both the content and the method of daily classroom activities are recognized. Wrinkle suggests six basic criteria for judging the value of each objective in a locally developed list of goals. "Is the objective (1) understandable, (2) stated as a behavior, (3) based upon the needs of the learner, (4) socially desirable, (5) achievable, and (6) measurable?"²

Teachers who are working cooperatively on a list of objectives for their subject area will want to clarify the meaning of each objective by stating it in terms of changes in student behavior. For examples of objectives stated in behavioral terms, the reader is referred to Table 11.1.

After objectives have been stated in behavioral terms, they should be *organized* under a few major headings. The grouping of interrelated objectives in this way assists in the planning of an evaluation program. For example, if all the objectives involving interests are grouped together, it may be possible to develop an interest inventory that will obtain evidence on a number of related objectives. Similarly, if the objectives relating to study skills are grouped together, a test of study skills can be selected or developed that will measure student knowledge and skill in a number of these objectives.

² William L. Wrinkle, *Improving Marking and Reporting Practices* (New York: Holt, Rinehart and Winston, Inc., 1947), p. 97.

Survey of Techniques and Instruments

Since effective measuring instruments are difficult to construct, it is advisable to use published test materials whenever they provide valid data for the purpose and group for which they are being used. In measuring the scholastic aptitude of students, published intelligence tests are indispensable. Published achievement tests of the survey type are necessary to determine the present status of individuals or groups or to measure their progress over a period of a year or more. Published diagnostic tests aid the teacher in identifying students' specific retraining needs, although the use of observation and interview often increases the validity of diagnosis.

In most subjects, however, teacher-made tests must be relied on for measurement of short-term growth. In such tests, most questions relate to the learnings recently emphasized in the local instructional program. Moreover, for many significant objectives of instruction, no published tests are now available. For some of these objectives, the teacher may always have to rely on such informal and relatively subjective means as observation, interview, and self-rating techniques. Hence, his conclusions must be tentative unless data obtained by independent methods tend to converge or agree in supporting a generalization or conclusion.

The staff should develop a source list of possible techniques and instruments for each of the major objectives. As an illustration, such a list has been prepared for two major objectives.

OBJECTIVE	POSSIBLE TECHNIQUES AND INSTRUMENTS
The student uses effective study skills	<p><i>California Study Methods Survey</i></p> <p><i>California Tests in Social and Related Sciences, Elementary, subtests on "Reading of Maps" and "Knowledge of Geographic Terms"</i></p> <p><i>Iowa Tests of Basic Skills, Test W, Work-Study Skills</i></p> <p><i>Peabody Library Information Test</i></p> <p><i>Spitzer Study Skills Test</i></p> <p><i>Survey of Study Habits and Attitudes (Brown-Holtzman)</i></p> <p><i>Tyler-Kimber Study Skills Test</i></p> <p><i>Wrenn Study Habits Inventory</i></p> <p>Informal teacher-made exercises on the interpretation of maps, graphs, and tables</p> <p>Teacher-made tests of study skills</p> <p>Teachers' checklists for guided observation of students' work on a specific library assignment</p> <p>Instructional tests issued by publishers on the use of the dictionary, encyclopedia, and other reference books</p> <p>Self-rating checklists on study habits</p>

OBJECTIVE	POSSIBLE TECHNIQUES AND INSTRUMENTS
The student cooperates effectively in class activities	<p>Anecdotal records</p> <p>Systematic observation of students, using code numbers for notations on significant behaviors</p> <p>Observation of students' attitudes revealed in role-playing and dramatic play</p> <p>Recording evidence of students' participation and contributions in cooperative group enterprises, such as bringing relevant material to class</p> <p>Lists of criteria (developed by teacher and students) on the characteristics of a good class discussion and other skills of group work</p> <p>Self-rating charts based on such criteria</p> <p>Having students write out the endings for reaction stories concerned with cooperation in a group</p> <p>Informal teacher-made tests on knowledge of the skills of group work (for example, functions of the chairman and secretary of a committee, parliamentary procedures, and the like)</p> <p>Published tests, such as the <i>Behavior Preference Record</i>, or the subtest on "Understanding of Democracy" from the <i>California Tests in Social and Related Sciences</i></p> <p>Published or teacher-made attitude scales</p>

In the process of preparing such source lists, one should explore many possible techniques and instruments for evaluating growth toward each major goal of the educational program.

A school staff that has carried out the steps in planning outlined above has already laid the groundwork for the selection and development of evaluation instruments. As teachers have thought through in detail the meaning of each objective, they have partially developed bases for selecting the published instruments that meet their needs; they have partially formulated many criteria for evaluating learning products; they have developed the "raw material" for observation guides and rating scales on sportsmanship, cooperation, responsibility, and similar objectives.

Selecting Standardized Tests

Selecting valid tests is a major problem for the school. There are large numbers of published standardized tests from which to select but a relatively small number of usable tests with high validity for a specific purpose. It is generally desirable to use the same intelligence and achievement test batteries over a period of years so that comparable longitudinal data can be obtained and so that teachers will achieve familiarity with each test's

values and limitations. For this reason, great care in the selection of such tests pays dividends over a period of time.

Before a standardized test is adopted for school-wide or system-wide use, it may be advisable to try it with representative classes and analyze the results. Such an analysis may show that the test is too easy or too difficult, or may disclose problems with respect to adequacy of instructions, time required in scoring, or other factors that would affect the test's usability.

The selection of tests can never be a routine or entirely objective procedure. Professional judgment is needed in deciding the relative importance of certain features of a test to be used for a specific purpose. The basic concepts involved in the appraisal of tests were presented in Part One and summarized in Chapter 5.

GUIDANCE WORKERS AND PSYCHOLOGISTS AS RESOURCE PERSONS

Principals and teachers often feel overwhelmed by the many decisions that have to be made in setting up a good evaluation program or in improving one that has been only partially effective. Fortunately, an increasingly large number of schools can request assistance from persons with special competence in measurement and evaluation. Large school districts often have a specialist in measurement, employed in a department of research and guidance. In smaller school districts assistance may be available from the county school department or from a nearby college.

Guidance specialists (such as directors of guidance or counselors) have usually had considerably more training in child-study and evaluation techniques than most teachers and supervisors. If a guidance specialist is assigned time for this purpose, he can be very helpful in the processes of selecting tests, planning the testing schedules, instructing teachers in the many details of administration and scoring, and planning and supervising an efficient system of record-keeping.

The functions of school psychologists vary considerably from one school system to another. Like the guidance workers, the school psychologists are generally well versed in child study and evaluation and can serve as consultants in all phases of an evaluation program. Their special contribution is usually in the individual testing of students who deviate far from the average in intelligence, personality, achievement, or conduct.

The administrators, supervisors, and teachers in a system that employs competent guidance workers and psychologists are fortunate in being able to call upon them for help in many aspects of the evaluation program.

Without specialized personnel, the problems of evaluation are more difficult, but they can be solved by guidance-minded administrators and teachers with released time and special training.

PLANNING THE TESTING PROGRAM

The responsibilities of all school personnel (teachers, administrators, supervisors, and guidance workers) must be kept in mind in the planning of the testing program and the selection of tests.

Aptitude Testing

Ideally, group intelligence tests should be administered in alternate years beginning with grade 2 or 3 and continuing throughout the elementary school years. In the first grade, a reading-readiness test seems preferable to an intelligence test. Not only is a reading-readiness test valuable in helping assign children to reading groups but teachers are less likely to make premature generalizations about the *general* learning ability of young children than if an intelligence test is given.

For the typical school, tests of general mental ability, heavily loaded with verbal and numerical content, would seem to provide the best dividend for a given investment of testing and scoring time. Such tests provide a better basis for assessing expected achievement than do tests of equal length that include spatial and perceptual factors less closely related to achievement in school.

For students who are markedly retarded in reading, however, verbal intelligence tests may grossly underestimate the student's potential for school achievement, once his reading handicap has been removed. For such students, the use of a supplementary intelligence test that does not depend unduly on reading skill is necessary. Schools that enroll a large percentage of children who are bilingual or come from culturally deprived areas, may prefer to give nonverbal, as well as verbal, tests to all pupils.

At the junior high school level, ability tests that provide separate scores in verbal and numerical abilities and possibly other factors of mental ability are valuable in helping students make decisions concerning elective subjects and tentative vocational choices.

If the testing budget permits, and qualified staff members can provide leadership to students and teachers in a self-appraisal program, the administration of a multiscore intelligence test or an aptitude test battery is desirable in the ninth or tenth grade. A school staff, however, should not embark on such a program unless they are willing to allow enough testing

time to obtain reliable scores on each aptitude. A two-factor test, giving scores on verbal and numerical aptitudes only, is preferable to a battery giving scores on five or more abilities if the subtests of the battery are too short to provide reliable scores.

Prognostic tests in algebra, geometry, foreign language, and shorthand might be included to advantage in some testing programs. Adequate expectancy charts can usually be constructed on the basis of intelligence test data and previous grades in relevant subjects; however, prognostic tests may be preferred because of the greater ease with which *special* aptitudes, rather than general scholastic aptitude, can be discussed with students and parents.

Achievement Testing

A minimal achievement testing program in the primary grades (grades 1-3) would involve the testing of reading skills in grade 2 and the testing of reading, arithmetic, and language in grade 3. In the upper elementary grades, annual administration of a test battery in the basic skills is recommended. The use of separate answer sheets with upper grade students may make possible the inclusion of tests in work-study skills, social studies, and science at the fifth- and sixth-grade levels with only moderate additional cost.

At the secondary school level achievement testing is of two types: (1) the school-wide testing program in the basic skills, and in the skills and content learnings of required general-education courses; and (2) the continuous day-by-day evaluation and the end-of-course testing in specific subjects. The second aspect of the program has been considered in Chapters 10 through 13.

Emphasis on the first aspect of the program has increased in recent years as high schools have assumed greater responsibility for the developmental and corrective teaching of the basic skills, as teachers have learned to use test data on students' cumulative records, and as school staffs have felt the need for having periodic evaluation data, both for appraising their own instructional programs and for informing the public concerning their effectiveness. The development of machine-scoring techniques and other rapid-scoring devices has helped to reduce the cost, and increase the practicability, of large-scale testing programs at the high school level.

A basic program of high school achievement testing should probably include:

1. The administration of a basic-skills battery to all incoming seventh-grade students (preferably during the last month of the sixth grade so that the data would be available as an aid in the programming of students into seventh-grade classes).

2. The administration of a basic-skills battery to all incoming ninth- or tenth-grade students entering four- or three-year senior high schools, respectively.
3. Administration of a test battery to evaluate student achievement in the skills and content learnings of general-education courses.

The following batteries are typical of tests available for use in evaluating the outcomes of general-education courses at the secondary school level:

California Tests of Social and Related Sciences, Advanced (Parts I and II at the completion of American history and Part III at the completion of required courses in general science and biology)

Cooperative General Achievement Tests, grades 9-13

Essential High School Content Battery, grades 10-13

Iowa Tests of Educational Development, grades 9-12

Metropolitan Achievement Tests, Advanced, grades 7-9

Sequential Tests of Educational Progress, grades 7-12

In both the elementary and secondary schools, it is important that every student be tested with a test that is at approximately the right difficulty level for him so that there are a large number of items on which he can demonstrate his competency. In the Atlanta schools, students are assigned to take tests at their reading level.³ In other school systems, teachers' judgments regarding the level of test that should be administered to each student are used in assigning students to testing groups. If STEP or SCAT tests are used, different levels may be administered at the same time in the same classroom.

Testing of Interests and Personal-Social Adjustment

If vocational-interest inventories are used as part of a student self-appraisal program, they should be administered at the same grade level as the aptitude test battery so that the interest-inventory results may be interpreted in the light of those from aptitude tests. The hazards of interpreting vocational-interest tests without accompanying data on students' abilities have been emphasized in Chapter 7. Because vocational interests change and mature, and because one interest inventory can serve as a cross-check on another, many students may desire to take a second vocational-interest inventory in the eleventh or twelfth grade as part of a vocational-planning unit or of an elective course for students who need additional help in life planning.

Tests of personal-social adjustment should be used only in situations

³ Warren G. Findley, "Use and Interpretation of Achievement Tests in Relation to Validity," *National Council on Measurement in Education* (Ames, Iowa: The Council, 1961), pp. 23-34.

in which the results are likely to be wisely interpreted and when psychological service is available to help individual students who need such assistance. In the hands of qualified staff members, personality inventories can be used to advantage, especially in identifying (with the help of teacher observation and other techniques) individual students who should be referred for special study.

Youth-problem checklists probably involve less risk of misinterpretation and are more meaningful to counselors, teachers, and students than the type of inventory that yields a profile of personality components or traits. Insufficient research has been done on the construct validity of most personality inventories. The subtests on which the profile is based may be so unreliable that a change in the student's answers to only two or three questions may cause a marked shift in percentile rank. The danger that too much may be "read into" such a personality profile by the naïve user should not be dismissed lightly.

Supplementary Testing

Up to this point, we have considered tests to be given to *all* pupils. Supplementary testing is essentially of three types:

1. Diagnostic testing of students tentatively selected for remedial work in the basic skills of reading, arithmetic, and language.
2. The administration of special interest and aptitude tests to students who need additional help with problems of educational and vocational guidance (for example, a group test of engineering aptitude to seniors considering a career in engineering, individual tests of manual dexterity to students considering careers as dentists or dental technicians, or the Strong interest inventory and a number of aptitude tests to a handicapped student or any other student facing special problems of vocational choice).
3. The administration of individual intelligence tests, personality inventories, and personality tests of the projective type to students referred to the counselor or psychologist because of marked and persistent underachievement or serious problems of personal-social adjustment.

For those students whose group-intelligence-test results are markedly inconsistent, individual intelligence tests are desirable. In addition, all students with group-test IQ's of 75 or below should be tested individually, as well as all those whose physical or emotional handicaps would tend to invalidate group-test results.

Scheduling of Tests during School Year

The basic or school-wide testing program is usually scheduled for either the beginning or the end of the school year. One advantage of a consistent

policy in this respect is that a year (or a multiple thereof) intervenes between successive testings. Hence, data on student growth are more easily interpreted from the records.

Fall, or beginning-of-year, testing has certain clear-cut advantages: (1) It permits the teacher to obtain a complete test record for each student. When students have been tested the preceding spring, pickup testing is necessary for new entrants. (2) The data are up-to-date. During a long vacation, many students lose in varying degrees their proficiency in certain skills; on the other hand some students have gained in reading achievement through their summer reading. Others may have gained in skill subjects through attendance at summer school or through special tutoring. (3) Fall testing places the emphasis on the analysis of student needs, rather than the evaluation of teaching. (4) More time is available for the administration and scoring of tests and the analysis of results. End-of-year pressures can result in tests being filed away without being used. (5) Up-to-date test results can be used as a basis for grouping students for differentiated work or special corrective instruction. Moreover, scores on survey tests serve as a starting point for the use of supplementary diagnostic methods to determine specific retraining needs.

End-of-year testing also has certain advantages. If tests are administered during the last month of school, teachers and counselors will have (1) recent objective data to aid in problems of promotion, and in programming students for the next year and (2) data that can be studied by teachers in the fall (either before or just after the opening of school). Some school districts have a systematic pre-session program that includes a study by the teachers of the cumulative records of incoming students. Obviously, testing of achievement in specific high school subjects involves end-of-year testing.

ADMINISTERING THE TESTING PROGRAM

Although many teachers and other staff members are involved in the planning and carrying out of a testing program, the responsibility for the basic program of standardized testing should be centralized. That is, one well-trained staff member at the central office should be working with one well-trained person at each school who will see that the many administrative details involved in giving and scoring tests are adequately observed. Tests and answer sheets need to be ordered well in advance of testing; test materials need to be distributed to teachers well in advance of use; testing schedules need to be planned and materials routed; examiners and proctors need to be trained in the administration of an unfamiliar test; supplementary directions regarding scoring may need to be prepared, and a sampling

of work at each stage of the scoring process must be checked. Unless testing is done under standard conditions and tests are accurately scored, the results will be inaccurate and misleading. It is imperative that responsibility for administering the testing program be centralized in a person with training, experience, and a realization of its importance.

The responsible staff member will probably find it desirable to prepare a bulletin covering all pertinent aspects of the school testing program and an abbreviated list of procedures *for each test* with references to appropriate sections in the test manuals. Such a list can be affixed to each package of test materials checked out to individual teachers.

Administration of Standardized Tests

It is essential in the administration of standardized tests that students be tested under *standard conditions*—that is, with the same instructions and the same timing as were the students on whom the test was normed. Unless the testing conditions are the same, the norms cannot be considered applicable.

PREPARATION FOR TESTING If valid results are to be achieved, the examiner must be thoroughly familiar with the test instructions, must have all testing materials at hand, and must have made arrangements so that the testing session will not be interrupted. The importance of advance preparation cannot be overemphasized. In group testing, there must be no emergencies.

The examiner should have so thoroughly familiarized himself with the instructions that he can read them with clarity and give proper emphasis to key words and phrases. He must have rehearsed every step of the process, knowing when and how he must demonstrate a sample exercise, when students should be asked to read the directions with him, and the like. If a test is timed, the examiner will need a watch with a second hand (or preferably a stop watch, the use of which will give him more freedom for the observation of students). The watch should be checked to see that it is operating properly.

Before he assembles his testing materials, the examiner should prepare a list of items needed. This may include scratch paper for certain parts of the test, extra pencils and erasers, and sometimes special supplies for machine scoring. Once the materials are assembled, they should be counted out by rows (that is, the number of tests, answer sheets, and the like equal to the number of seats in row 1, row 2, and the like).

If machine-scored tests are used, pencils should be checked to make sure that they are sharpened and that the erasers are in good condition. If mechanical pencils are used, they should be checked to be sure that each pencil contains sufficient electrographic lead.

SCHEDULING OF TESTS Before administering a test, the teacher should make sure that the students can complete the test before the end of the period. Freedom from distractions is very important. He should protect the students from interruption by placing a sign on the door reading "Testing—No Admittance," and should inform the office that testing is planned so that there will be no interruptions through the interoffice telephone.

Rapport is extremely difficult to establish if a test is given late in the day, on the day before vacation, or during a period of undue excitement (such as before a school athletic event). Hence, these periods should be avoided in setting up a test schedule.

With elementary school pupils, one should avoid having the test extend into the recess period. An opportunity to visit rest rooms and get a drink should be provided before the test begins. If only part of a class can be tested at one time, arrangements should be made for the supervision of the other pupils, outside the classroom. Third-grade classes, as well as some second-grade classes, can be tested as a group; however, the teacher may need the assistance of another staff member to serve as proctor, making sure that pupils are following directions, working on the right page, and continuing to work throughout the testing period.

In planning testing schedules, adequate time should be allowed for distributing papers, giving instructions, answering preliminary questions, and, following the test, for collection of tests, answer sheets, pencils, and other testing materials. It is also important not to expect students, especially younger ones, to work too long in one testing session. Usually the manual for a test battery will suggest how the testing time should be divided into sessions. It is better to spread testing over several days than to have children become bored or frustrated.

If students' answers are recorded on answer sheets rather than test booklets, the school may route classroom sets of test booklets from class to class, rather than purchasing booklets for all examinees. If sharing of supplies is involved, the testing must not be too closely scheduled. When classroom sets are routed from class to class, the testing schedule should allow enough time between testing sessions to allow for examining all booklets and screening out those on which students have recorded answers.

A library or cafeteria used for testing large numbers of students should be well lighted and well ventilated, and should have satisfactory acoustics. The use of tablet arm chairs is not recommended, especially if separate answer sheets are to be used.⁴ Seating arrangements should be planned so as to provide good working space and minimize cheating. The examiner may wish to assign to front seats children who are likely to have difficulty in following directions.

⁴ Arthur E. Traxler and R. N. Hilkert, "Effect of Type of Desk on Results of Machine-scored Tests," *School and Society*, vol. 56 (September 1942), pp. 277-279.

If the time limit for the test is liberal, the teacher should plan in advance the best procedure to follow with students who finish early. They should be directed to work on some activity that will not require them to move about the room and that will not require supervision by the teacher. Instructions regarding permissible activities should be given before the test begins. If a series of tests is being given to a large group, the last test should be a closely timed one so that all students will complete their work at the same time.

Questions *after* testing has begun should have been discouraged by the examiner's initial presentation and should have been made unnecessary by the clarity of his instructions and examples. If an individual student is perplexed about procedures, a proctor can repeat original instructions; but if a student asks for help on a *specific item*, he must be told, "I'm sorry that I cannot answer your question. Do your best. If you are stuck, go on to the next question."

Communication and Rapport

One of the most difficult aspects of test administration for the teacher is that of "doing enough but not too much" in helping students to understand the test. The teacher's role as an impersonal examiner is an unnatural one for him. In his eagerness to help students understand the test, the teacher may be tempted to reword the directions. Such rewording not only violates the requirement of standard conditions but also may reduce the effectiveness of communication to the students. The instructions of a well-constructed test have been tried out and revised a number of times before publication.

If a large number of students are being tested, a sufficient number of proctors should be assigned—in general, 1 for every 20 to 25 students. Proctors work most effectively when they have carefully studied the test and the test directions.

As the students are taking the test, the teacher (and proctors) should move about the room unobtrusively to make sure that each student is recording his answers in the proper way and that he is working on the right part of the test. The teacher should avoid moving quickly about the room, watching a student over his shoulder, carrying on a conversation with a proctor, or doing anything else that will distract the students. The teacher should make notations concerning any student behavior that might affect the test results, such as undue anxiety, distractibility, dawdling, or needing to leave the room.

Interpretation of results from ability tests rests on the assumption that each student has done his best. Hence, every student should be motivated to make his maximum effort. The motivating talk should be brief and to the point. Usually it is included in the test manual. At least two important

factors affect student motivation in taking tests: (1) the sense of inadequacy that many students experience when they confront items they cannot answer, and (2) the sense of indifference that many students exhibit when they are confronted with any arduous task that has little meaning for them. The test administrator can meet the first problem by explaining that the test is designed to measure achievement over a wide range of content and difficulty, and will include problems based on material that students have not been taught. If the test is closely timed, the examiner can also explain that many will not be able to finish their work. The problem of indifference can be handled largely through the examiner's attitude of alertness and interest in administering the test. An effective means of combating student indifference is to indicate to the students how the test results will be used to help them.

A comprehensive, well-organized checklist for persons who give standardized tests has been prepared by Thompson.⁵ The items in the checklist are conveniently organized into four sections: (1) before tests are given, (2) during the testing, (3) after the testing period and (4) at all times.

Since standard conditions are so important, a number of schools have tried having a well-trained examiner dictate test instructions over the intercommunication system to the rooms involved. Such a plan, however, is so inflexible that there is no opportunity to adapt to any emergencies that might interfere with the progress of a single class. The Los Angeles schools have developed a phonograph record of test instructions, to be used with the *Kuhlmann-Anderson Intelligence Test*.⁶ The use of a record would provide for somewhat more flexibility.

The use of sound films, filmstrips, and tapes in training teachers to administer tests, and the certifying of teachers who have taken short courses on the administration of group achievement tests (or a *specific group intelligence test*) represent more desirable steps toward approaching standard conditions of administration than the use of impersonal methods. For example, a color filmstrip with a synchronized long-playing record is concerned with the administration, scoring, and interpretation of the *Iowa Tests of Basic Skills*.

Scoring of Standardized Tests

The care taken in the selection and proper administration of tests is to no avail if the tests are inaccurately scored. Errors are of two types: (1)

⁵ Anton Thompson, "Test-Giver's Self Inventory," *Test Service Bulletin* No. 85 (New York: Harcourt, Brace & World, Inc., n.d.). Copies available on request.

⁶ Howard A. Bowman, "Assisting Teachers in Test Administration," *19th Year-book*, National Council on Measurement in Education (Ames, Iowa: The Council, 1962), pp. 61-63.

constant errors, resulting in scores that are consistently too high or too low, due to failure to understand the scoring instructions; and (2) *variable* errors, caused by carelessness in marking, computing, or copying scores.

Whenever a teacher is scoring a standardized test for the first time, he should score three or four tests and ask to have the scoring checked. Such an early check-up may avoid the repetition of errors due to a misunderstanding of instructions. Even teachers who are familiar with the scoring instructions should have a sampling of tests rescored. If careless errors are discovered, all tests for that class should be rescored.

Accuracy and speed of scoring are usually improved if the teacher scores test 1 for all students, test 2 for all students, and the like, instead of scoring all the tests in each student's booklet before scoring the next student's booklet. In this way, he handles only one scoring key at a time and can keep in mind any special instructions for scoring the test. No matter how familiar the key becomes, the teacher should never score from memory.

A number of devices have been developed to reduce the burden of test scoring. An outstanding development has been the invention of the electrical test-scoring machine.⁷ This equipment, now available in most city and many county school systems, requires the use of machine-scored editions of the tests and of special answer sheets. Schools not having machine-scoring equipment can hand score the separate answer sheets by means of scoring stencils with holes in the correct answer positions. Many test publishers provide machine-scoring service on their own tests on a fee basis.

A number of publishers have given considerable attention to the designing of tests and keys so as to facilitate hand scoring. For example, equated scores or grade placements are often printed adjacent to the raw score equivalents on the test booklets or answer sheets. Test publishers have also prepared tables and other devices to assist in the computation of chronological ages from birth dates and intelligence quotients from data on raw score and age. When a computational aid that increases the speed of test scoring is used, spot checking is necessary to make sure that it is being used correctly.

SUMMARY STATEMENT

The functions to be served by a school-wide evaluation program may be classified as (1) administrative-supervisory, (2) instructional, and (3) guidance functions. In order to serve all these functions effectively, an evaluation pro-

⁷ International Test Scoring Machine (International Business Machines, 590 Madison Ave., New York, 22, N. Y.). Procedures to be followed in achieving specified degrees of accuracy in machine scoring are detailed in a manual that may be obtained from the company: *IBM Test Scoring Manual of Procedures*.

gram should be comprehensive, continuous, and flexible. It should be developed through cooperative planning, be based on the objectives of the educational program, and keyed closely to the local situation. Such a program necessarily involves a variety of evaluation techniques.

In planning a school evaluation program, the teaching staff needs to make an analysis of the major objectives of the educational program and to survey possible techniques and instruments which can be selected or developed to measure student progress toward these objectives. In selecting standardized tests that can be used to advantage, the relevant criteria formulated in Chapter 5 need to be applied.

General recommendations were made concerning the planning of different aspects of the school testing program: the selection, scheduling and administration of aptitude, achievement, interest and personality tests. Needs for supplementary testing for individuals and for specific subgroups should also be considered.

For effective administration of school testing programs, responsibility needs to be centralized (both at the school and school system levels) in persons who are well trained for such responsibility. Proper precautions need to be taken to make sure that persons administering tests are adequately prepared and adhere to standard instructions; that tests are administered under optimum physical conditions; that communication and rapport are sufficiently good that examinees are well motivated; and that the scoring of tests is accurately done.

SELECTED REFERENCES

- CROOK, FRANCES E., "Elementary School Testing Programs: Problems and Practices," *Teachers College Record*, vol. 61 (November 1959), pp. 76-85.
- GORDON, LEONARD V., "Right-Handed Answer Sheets and Left-Handed Testees," *Educational and Psychological Measurement*, vol. 18 (Winter 1958), pp. 783-785.
- PHILLIPS, BEEMAN N., AND GARRETT R. WEATHERS, "Analysis of Errors Made in Scoring Standardized Tests," *Educational and Psychological Measurement*, vol. 18 (Autumn 1958), pp. 563-567.
- SARASON, SEYMOUR B., "What Research Says about Test Anxiety in Elementary School Children," *National Education Association Journal*, vol. 48 (November 1959), pp. 26-27.
- , AND OTHERS, "A Test Anxiety Scale for Children," *Child Development*, vol. 29 (March 1958), pp. 105-113.
- SMITH, WILLIAM F., AND FREDERICK C. ROCKETT, "Test Performance as a Function of Anxiety, Instructor and Instructions," *Journal of Educational Research*, vol. 52 (December 1958), pp. 138-141.
- SUPER, DONALD E., AND JOHN O. CRITES, *Appraising Vocational Fitness*. New York: Harper & Row, Publishers, Inc., 1962, Chapters 4, 5.
- THOMPSON, ANTON, "Tentative Guidelines for Proper and Improper Practices with Standardized Achievement Tests," *California Journal of Educational Research*, vol. 9 (September 1958), pp. 159-166.
- , "Test-Giver's Self-Inventory," *California Journal of Educational Research*, vol. 7 (March 1956), pp. 67-71.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. List the types of information you would like to have concerning students in a class you are to teach. To what extent could such information be obtained from standardized tests?
2. Outline a basic testing program for either an elementary or high school. Indicate the types of tests to be used and the frequency and time of administration.
3. Report on the uses of tests of scholastic aptitude in a specific school system. Under what circumstances are individual intelligence tests given to pupils?
4. List several major objectives for a subject area in which you plan to teach. Which of these objectives could be effectively measured by written achievement tests?
5. What are the advantages and disadvantages of having uniform state-wide testing programs in the basic skills? In the content subjects?
6. Why should the responsibility for test administration be centralized at the district level and also within the school?
7. In what ways have the electronic methods of test scoring affected the development of testing programs?

*Summarizing, Recording,
and Reporting Data
about Individual Students*

The summarizing and recording of data can be time-consuming processes of little significance. On the other hand, they can be carried on in such a way that new understandings of the student are achieved: (1) through putting together current data on various aspects of behavior and thus perceiving their interrelationships and (2) through building up a longitudinal picture of a student's development. The process of reporting to students and parents can become routine drudgery or can be a valuable and stimulating process involving a sharing of insights and cooperative planning. In order that all three processes (summarizing, recording, and reporting) can become as functional as possible, it is necessary to keep constantly in mind that they are means to ends, not ends in themselves.

SUMMARIZING AND RECORDING DATA

If measurement data are to be used to best advantage in achieving a better understanding of students and their problems, an adequate, functional *cumulative-record system* is necessary. The term "cumulative-record system" is used to include all forms and procedures involved in maintaining a continuous record of each student's growth and development.

In a school system that makes adequate provision for recording personnel data, the cumulative-record system ordinarily consists of (1) a comprehensive form (known as the cumulative-record form), which includes identifying data and information on the student's home environment, scholarship, test scores, attendance, health, and the like; (2) supplementary card forms containing more detailed information on the student's attendance, health examinations, and the like; and (3) a cumulative-record

folder¹ in which are filed such items as test booklets, anecdotal records, and reports on student and parent interviews.

Purposes of a Cumulative-Record System

The cumulative-record system serves many purposes for staff members who work with the student. The cumulative record provides the official record of a student's attendance, achievement, promotions, and graduation. It constitutes the basis for his transcript of record when he transfers from school to school. For incoming students, as well as those already enrolled, the cumulative records provide data that assist in determining each student's appropriate assignment to grade level and to specific classes.

The study of data recorded in a student's cumulative records can also help teachers to understand his behavior—to discover special needs, to distinguish between transient and more permanent behavior tendencies, to find out when a problem started, and to discover clues concerning causal factors underlying a student's difficulties. These purposes are achieved not only through the compilation of basic facts recorded on the cumulative-record card about the student's health, home situation, learning ability, and so forth but also through many informal reports filed in cumulative-record folders—observations of behavior, reports of student and parent interviews, and autobiographies and other self-expressive materials.

An adequate cumulative-record system also aids teachers and other school personnel in helping parents to achieve a more objective and accurate picture of the student's achievements, special abilities, and special problems. Kawin emphasizes the great potentialities of using the cumulative-record system to help parents to understand their children; yet she warns concerning the hazards of such a process.

Records can be invaluable in home-school contacts if the school personnel know how to use its records wisely and constructively in dealing with parents. No school device is more effective when wisely used, but no material is more likely to antagonize parents if wrongly used. On the whole, no school record should just be handed to parents for them to look over. Selected parts of records should be shown to parents by a member of the school staff who is competent to interpret this material constructively.²

¹ In a large number of school systems, the cumulative-record form is printed on a folder in which tests, observational records, and the like can be filed. Thus item 1 not only serves to provide a summary of the types of data listed but also fulfills the function mentioned in item 3.

² Ethel Kawin, "Records and Reports; Observations, Tests, and Measurements," *Early Childhood Education*, 46th Yearbook, National Society for the Study of Education (Chicago: University of Chicago Press, 1947), p. 290.

Characteristics of a Cumulative-Record System

Although the specific items desired may vary from one school district to another, there are certain general characteristics that are essential in any good cumulative-record system.

1. A cumulative record (preferably record folder) should be started for each student at the time of his entrance to school.
2. The records should be transferred (at least in summary form) as the student progresses from lower to higher schools or moves to another school district.
3. The cumulative records should present as comprehensive a picture as is feasible of the student's growth and development. Provision should be made for the cumulation of both test and nontest data.
4. The forms used should be simple and easily understood. Their maintenance should require no more clerical work than can be justified by the use of data recorded.³
5. The cumulative-record system should be flexible, requiring a minimum of data for all students but permitting great latitude in the types of additional data that are cumulated for individual students. The use of a record folder permits such flexibility. For example, records of interviews with parents or previous teachers can be filed in such a folder.

Camden recommends that high school counselors tape-record interviews with eighth-grade teachers concerning the students who will soon enter the upper secondary school.⁴ Flexibility can also be provided by including in the cumulative-record form a sufficient amount of blank space for significant teacher comments and for summary statements. Still another aid to flexibility is the provision of space on the cumulative-record form for rating special interests and problems as well as for notations on the location of significant information that is too voluminous or too confidential for entry on the record form.

6. The cumulative-record system should be designed to reveal trends in growth over a period of years. This criterion implies the recording of as much objective data as possible. It implies the use of comparable tests of scholastic

³ The rapid progress being made in machine processing of educational data may soon result in widespread changes in the posting of data on secondary school cumulative records. In 1959 the Educational Testing Service, working cooperatively with school and college educators in Georgia, developed a pilot project in the maintenance of student records, organizing and summarizing data about students with the aid of electronic equipment, and reporting such data on comprehensive student report forms. Since the plan is a flexible one that can be adapted to various needs, it attracted the attention of educators in other states. A grant from the Ford Foundation early in 1962 enabled the Educational Testing Service to work with educators in seven other states in the development of pilot projects similar to the one in Georgia. The cooperative Plan for Guidance and Admission now has a national advisory committee. Five regional conferences were held during the 1962-1963 school year to explore the further development of the plan. *Annual Report, 1961-62* (Princeton, N. J.: Educational Testing Service, 1962), p. 55.

⁴ Blanche Camden, "For a Better Understanding of Entering Students," *The School Review*, vol. 61 (January 1953), pp. 40-42.

aptitude and achievement as well as comparable measurements of height, weight, and other characteristics. Moreover, the record form should be so designed that data that are cumulative can be presented in chronological sequence. All entries should be dated. Data regarding the student's personal-social adjustment (gleaned through observation of behavior, conferences, study of creative writing, and the like) should be *summarized* at the end of each school year to reveal evidences of growth, as well as special needs and problems. At the time such a summary is made, much of the supporting data may be discarded. However, unusually significant materials (such as case studies, records of parent conferences, student autobiographies, and profiles of recent tests) should be retained in their original form.

7. Cumulative records should be readily accessible to teachers. However, the confidential nature of the data must be respected, and the records always kept in locked files. Some material in a student's record may be so highly confidential that it should be kept in a special file directly accessible only to administrative and guidance personnel. In such circumstances the entry "See confidential file" can direct the teacher to the administrator or guidance worker for an interpretation of the material.
8. Cumulative records should be so maintained that the data are accurate, complete, and up to date. Data on a student's test results are incomplete unless the following information is given:
 1. Complete name and identification of test—for example, *New Standard Achievement Test, Form A, Intermediate Level*
 2. Date on which the test was administered—for example, January 12, 1964
 3. By whom the test was administered
 4. Type of score. In all cases in which a standardized type of score has been adopted, abbreviations are adequate to identify the type of score.⁵
9. In the recording of data, every attempt should be made to distinguish facts from personal opinions. As cumulative records are expanded to include data obtained through informal evaluation techniques, it is important that teachers distinguish between objective facts and subjective impressions. The importance of making this distinction in anecdotal records was emphasized in Chapter 8.

If the data in cumulative-record folders are to be of optimal value, they should be (1) organized for ready accessibility, and (2) culled out periodically, the older and less valuable material being discarded after a summary is made. As a basis for organizing materials in the record folder, the following plan is suggested:

1. Prepare and mimeograph for school-wide use a simple summary sheet to be stapled inside each student's folder. As a basis for summarizing entries in the record folders, a grid similar to the one in Figure 16.1 is desirable. Areas to be listed in the left-hand column can be established through group planning.

⁵ Adapted from *Handbook of Cumulative Records*, A Report of the National Committee on Cumulative Records, United States Office of Education, Bulletin 1944, No. 5 (Washington, D. C.: Government Printing Office, 1944), p. 43.

Area of Information	GRADE					
	7	8	9	10	11	12
1. Intelligence						
2. Health						
3. Personal-Social Adjustment						
4. Environment						
5. Achievement						
5A. Basic Skills						
5B. Content Subjects						
5C. Work-Study Habits and Skills						
6. Special Interests and Talents						
7. Social Attitudes						

Fig. 16.1 Suggested Summary Sheet for a Cumulative-Record Folder.

2. Number serially all materials filed in a student's folder at the time they are filed (that is, the first material filed would be entry 1; the next, entry 2; and the like.)⁶ These numbers should be written in red at the upper-right-hand corner of each test booklet or other item filed.
3. At the time an entry is filed and numbered, its number should be entered for the proper grade level opposite the "Area of Information" to which it pertains. In some cases, more than one such entry may be made. A report of a home interview, for example, might be entered not only under area 4 but under other areas in which significant information was provided (for example, area 3 or 6). A report card covering all aspects of a student's achievement would be listed under area 5, whereas a test limited to the basic skills would be entered under 5a. An anecdotal record might reveal significant information in both areas 3 and 7.
4. If the evidence filed reveals a special problem, the number could be encircled.

The plan suggested above is only illustrative. The summary sheet used and the symbols employed should be developed locally and modified in the light of local experience.

Time should be provided for elementary school teachers, and homeroom or guidance teachers at the high school level, to make a careful study of

⁶ If material is temporarily taken out of a student's folder, a blank sheet should be inserted in its place (entry 14—Miss Smith—date) so that anyone using the folder would know where the missing material can be found.

the cumulative-record folders of incoming students before school opens. If such time is not provided, teachers should make such a systematic study immediately after the opening of school. Such early study will enable the teacher to obtain a picture of the average level and the variability of his class with respect to scholastic aptitude and various aspects of achievement. It will also enable him to identify students in need of corrective instruction or other individual attention. On the basis of these data, the teacher will be able to plan for needed observational records and for student and parent interviews so as to obtain additional data early in the year for those who need assistance with special problems.

Procedures should be developed for disseminating to high school teachers data about the aptitudes of all their students and their achievement in the basic and work-study skills. Relevant data from standardized tests, as well as significant nontest data, could be distributed through the student if the data are reproduced on punch cards and not "interpreted" or printed. That is, each high school student might receive at registration time a pack of six or seven data cards (on which data were punched but not printed). The teacher of each class would collect cards from students each period. Each teacher could then turn in the cards for all of his students grouped by period. The teacher's code number and the appropriate period number could be gang-punched on the cards. Then, the names of students, accompanied by significant test and nontest data, could easily be listed by machine on roll-book sheets for each teacher's classes.

REPORTING DATA TO STUDENTS AND PARENTS

Few issues have occasioned more discussion among parents and teachers during the past decade than those concerned with assignment of marks, report cards, and the entire process of reporting to parents. Much of the controversy has probably grown out of differences in the emphasis placed on various functions of the report card by parents, teachers, guidance workers, and administrators.

Functions of a Program of Reporting

The major functions of any reporting plan include the following:

1. *Administrative functions*—to provide data for use in promotion, transfer, and graduation.
2. *Guidance functions*—to identify areas of special strength and weakness as a basis for realistic self-appraisal and future educational and vocational planning.
3. *Motivational functions*—to stimulate students to increased effort in order to earn good marks.

4. *Informational functions*—to inform the student and his parents concerning his progress toward the goals of the educational program as a basis for co-operative planning.

Teachers and administrators who place great emphasis on the first two functions tend to favor letter or number grades, which are easily recorded and have the appearance of being objective and comparable. These educators tend also to favor marking a student in terms of his relative status in comparison with other students. Unless this is done, they believe, the cumulative record of teacher's marks will provide no accurate basis for knowing what the student's achievement has been or whether he should take certain high school subjects, such as advanced mathematics and science.

Of the functions listed above, it is the first one that seems to be best served by the traditional report card. A single letter or number does provide a convenient, easily recorded symbol of the teacher's judgment concerning a student's work. Such symbols can be used to compute grade-point-averages and rank in graduating class. A cumulated record of such marks can be photographed to provide transcripts for other schools.

Use of students' marks to serve the second function (that is, to identify special strengths and weaknesses as a basis for educational and vocational planning) involves the assumption that grades are comparable. The way in which grading standards vary from teacher to teacher minimizes the value of grades for this second function. However, over-all high school grade average does have predictive value for academic success in college. In fact, high school grade-point-average tends to correlate as high with college grades as student scores on a long achievement test battery (such as the *Iowa Tests of Educational Development*). Moreover, student grades provide the only reported evidence of students' strengths and limitations in many subject areas in which standardized tests are not available, such as art, music, dramatics, and other fields.

Many parents, and also a number of teachers, stress the motivational function of marking and reporting. They tend to emphasize the need for a competitive marking system, which requires that a student be marked in terms of how his achievement compares with that of his classmates. On the other hand, many educators stress the disadvantages of competitive marking, for example, the discouragement of the dull student who does his best, the stimulus to cheating and to superficial achievement rather than genuine growth, the effects on parent-child relationships, and the like.

Many educators believe also that dependence should not be placed on the motivating value of marks—that "a course in which the mark is the major stimulus for the student to work should be discarded or subjected to extensive revision."⁷ Immediate information concerning specific successes

⁷ William A. Wrinkle, *Improving Marking and Reporting Practices* (New York: Holt, Rinehart and Winston, Inc., 1947).

or failures, such as can be provided through programmed instruction and evaluation of projects completed, may provide a sounder basis for motivation. If the student has many short tests or obtains (through these tests and through other means) many cues concerning his successes and failures, periodic summary marks may be unnecessary from the motivational point of view.

Elsbree points out that the fourth function—that of informing students and parents—is the primary function of any reporting system and that all other functions are incidental to it.⁸ A single letter or number grade is especially inadequate for the fourth function, that of providing information on each student's progress as a basis for cooperative planning with students and parents. Most of the modifications that have been introduced into school reporting systems have been designed to improve their value in communication.

Types of Reporting Practices

Wrinkle, who has experimented for more than a decade with different reporting practices at the high school level, states that departures from the traditional marking system (percentage or letter marks) consist either of (1) *manipulating the symbols* (for example, by changing to such symbols as "S" and "N" for "Satisfactory," "Needs improvement") or (2) *supplementing the symbols*.⁹ One purpose of the first approach, which is widely used in the primary grades, is to reduce emphasis on competitiveness among students and to reduce parental pressures upon the child. Such adaptations, however, communicate *less* information about the child. If they are supplemented by individual parent-teacher conferences, however, they provide a satisfactory solution for younger children.

The second approach has involved three techniques: (1) the development of fairly detailed rating scales of major and minor objectives on which the teacher rates the student, (2) the use of informal letters, and (3) the parent-teacher conference.

THE RATING SCALE Through the use of rating scales, teachers can give parents a more detailed picture of what the school is trying to accomplish in each major area and the student's relative strengths and weaknesses within each area. Since even the simplest rating scale includes a number of items, the teacher can be expected only to make gross distinctions with respect to student performance on each item. The rating scale has the advantage of being a simple way of reporting a great deal of information with a minimum of time and effort. The data are also easily recorded in permanent form on the student's cumulative record.

⁸ Willard S. Elsbree, *Pupil Progress in the Elementary School*. (New York: Bureau of Publications, Teachers College, Columbia University, 1943), pp. 72-73.

⁹ Wrinkle, *op. cit.*, p. 50.

The effectiveness of this type of reporting depends upon (1) the care with which the list of objectives is developed and the objectives are defined, (2) the extent to which parents and students are involved in developing the statement of objectives, (3) the adequacy of the teacher's techniques for evaluating progress toward each goal, (4) the care the teacher uses in the actual rating process, and (5) the extent to which students are involved in self-appraisal of their own growth and discussion of their own ratings with the teacher. A rating scale should be restricted to those aspects of student performance on which the teacher can reasonably be expected to have adequate evidence.

The following excerpts from the Pasadena City Schools Progress Report, grades 1 and 2, illustrate the way in which a major heading (on which a grade of "O," "S," "N" is given) may be followed by a checklist of behaviors.

Pasadena City Schools Progress Report¹⁰

	2d report	3d report	4th report
READING	S	S +	S +
Reads with understanding		+	+
Shows interest in reading	+	+	+
Works out new words	—		
Reads with fluency	—		
HANDWRITING	S	S	S +
Forms letters correctly	+	+	+
Is neat	—	—	+
PHYSICAL EDUCATION	O	O	O
Is developing coordination	+	+	+
Participates in group games	+	+	+
Uses equipment properly			
Exhibits good sportsmanship	+	+	+
Demonstrates skill in rhythms	—		
WORK AND STUDY HABITS	S	S	S +
Makes good use of time			
Follows directions			
Makes satisfactory effort	+	+	
Works independently	—	—	+
Listens attentively			
Does neat work	—	—	+
Uses materials wisely			

¹⁰ These grade symbols are interpreted as follows: O—Outstanding, S—at grade level, N—below grade level. A plus or minus sign may be used after any subhead to indicate strength or weakness. Since the first report each year is a teacher-parent conference, the first entry on the report card is the 2d report. Reprinted with the permission of the Pasadena, California, City Schools.

THE INFORMAL LETTER The informal letter has many advantages as a medium for reporting to parents. The letter can be individualized to highlight the special strengths and needs of an individual student. It can be highly analytical in those areas of the student's development in which specific problems are being met. A carbon copy of the letter constitutes a permanent record that should be filed for use by later teachers. The use of the letter form stimulates many parents to reply.

The informal letter, however, has certain limitations. Many teachers do not express themselves easily and effectively in writing. Even with teachers who do write well, there is greater possibility of misinterpretation by the parent than in a conference. It is difficult to report student weaknesses tactfully without minimizing them unduly. Because of these difficulties, informal letters frequently deteriorate into stereotyped reports of little interest and value. When this deterioration occurs, the values of the letter report fail to justify the large amount of time required.

THE TEACHER-PARENT CONFERENCE There is little disagreement on the value of the teacher-parent conference as a technique for communication with parents. Like the letter, the conference can be individualized to focus attention on those aspects of achievement and those problems that are most important to the individual. Through a conference, a variety of data and their interrelationships can be interpreted. The possibilities of misunderstanding are diminished. The parent has the opportunity to present his questions and problems. The teacher obtains information of value concerning the student; and, perhaps most important, a good conference leads to cooperative planning by teachers and parents.

An effective program of teacher-parent conferences requires teacher education concerning effective preparation and interview techniques, as well as some released time for conferences and for making adequate written records for future use. Many school districts operate on a shorter school day during two or three weeks of the time that conferences are scheduled.

Certainly the conference method cannot constitute the sole method of reporting unless adequate written reports are made. Even if such reports are made, however, there still remains the problem of marks or grades for students transferring to other schools, as well as the highly significant problem of reporting to the student. At present, almost all schools retain some type of letter-grade reporting at intervals throughout the year. These intervals vary from six weeks to an entire semester.

In recent years, school districts have been experimenting with the interpretation of test data to parents through these conferences. In fact, legislation and court rulings in some states (for example, California and New York) have required that parents be given information concerning the results of standardized tests if they so request. Since educators are rightfully concerned about misinterpretation of such data and the possibility

of harmful pressures on children, considerable attention is now being given to developing optimum procedures for the interpretation of test results to parents. In publications concerned with this problem, a number of helpful suggestions have been made:

1. One should avoid communicating intelligence quotients or other numerical scores; it is better to use an expectancy chart (or a verbal interpretation of one), to interpret the score in terms of its relationships to other significant variables. Intelligence should be interpreted as developed scholastic aptitude, rather than innate mental ability.
2. If scores are requested, probably stanine scores are the most suitable because they minimize the risk of the parent's attaching too much significance to small differences in raw scores.
3. The use of grade equivalents is to be avoided because of (a) their inadequacies in showing relative strengths and weaknesses (as explained in Chapter 2) and (b) the faulty inferences likely to be drawn about the student's competency to do work of a higher grade level.
4. The use of some device, such as the percentile band, which reminds the reader of the error of measurement, is desirable.

Ebel indicates one of the problems faced when IQ's are reported to parents.

One principal in a midwestern school ran into trouble with the parents of a 10-year-old who had an IQ of 110. A neighbor's daughter, who was only eight, had a measured IQ of better than 130.

"What do you mean," the irate parents demanded. "Our son is smarter than that eight-year-old. He's a better speller. He can do long division and fractions and she can't. But she got a higher score than he did. If she's so smart, why isn't she in the fifth grade, too?"¹¹

In the Winchester (Massachusetts) schools, *stanine* scores for as many as eight subtests of an achievement battery can be graphed on a form like that in Figure 16.2. The grid is nine blocks wide to correspond with the nine segments of the stanine scale. No explanation of stanine is printed on the chart and no numbers are used. A series of charts are used so that the teacher can select for each child the one that has the right shading for "Ability." In the example, the child's stanine for scholastic aptitude is 4; hence a chart is chosen on which blocks 3, 4, and 5 are shaded. In recording the data for each subtest, the teacher draws a red line starting at the left and going through the block representing the stanine equivalent of the student's raw score on that test. Because of the error of measurement involved, a student is considered to be achieving reasonably well if his achievement stanine on each test falls within the shaded area.

¹¹ Robert L. Ebel, "How to Explain Standardized Test Scores to Your Parents," *School Management*, vol. 5 (March 1961), pp. 61-64.

SCHOLASTIC ABILITY — ACHIEVEMENT CHART									
AREA TESTED			LOW		AVERAGE		HIGH		
Ability									
A	English								
C	Mathematics								
H	Science								
I	Social Studies								
E	Language								
V									
E									
M									
E									
N									
T									

Fig. 16.2 Form Used for Reporting Test Data to Parents in the Winchester (Massachusetts) Public Schools.

From Norton E. Demsey, Jr., "Reporting Test Results to Parents," *The 19th Yearbook, National Council on Measurement in Education* (Ames, Iowa: The Council, 1962), pp. 64-66.

Although the Winchester schools send this report home with the last regular report card, the advantages of using such a form at a teacher-parent conference in the fall should be seriously considered. On the whole, discussion of the chart at the conference and refileing it in the student's folder seems advisable, rather than sending it home, possibly to be misinterpreted to the other parent and to the child.

The following caution regarding interpretation should be communicated to parents either orally or in writing when test data are examined in a teacher-parent conference.

Parents should understand that the testing process is subject to a certain margin of error, and that the relationships depicted thereon are based on *one test of ability* as compared with *one test of achievement* in each of the areas listed. Thus, the picture shown cannot be considered precise and it is intended only as a guide to the child's recent achievement in relation to his potential.¹²

It may be advisable to average all available stanine scores from scholastic aptitude tests and reading tests so as to obtain a more reliable index of each student's ability to do academic work.

¹²Norton E. Demsey, Jr., "Reporting Test Results to Parents," *19th Yearbook, National Council on Measurement in Education* (Ames, Iowa: The Council, 1962), p. 65.

In Chapter 8, techniques of interviewing were discussed. Some additional hints that apply more specifically to conferences regarding the interpretation of test results are presented by Sax:

1. Try to find out the parent's reactions to the child's progress in school by such questions as: "How are things progressing?" "How do you feel we can be of help?" A discussion of these questions provides the teacher with an opportunity to describe test results as they relate to parent concerns and involves the parent more actively in the study of the data.
2. Check the cumulative record folder for notations concerning which test results have already been presented to the parent. Such care will minimize the risk of needless repetition or of contradicting the picture previously presented by a colleague.
3. Where the evidence appears contradictory or otherwise inadequate, be willing to admit the inadequacies of tests and discuss the reasons for variations in results.
4. Try to gear the type, amount, and complexity of the material presented to the parent's apparent ability to understand and utilize the results. Stress those aspects of the data which seem most relevant to "next steps" for the student in terms of remedial work, or other enrichment plans. Avoid technical terms, but do not over-simplify or "talk down" to parents.
5. Before closing the interview, it may be advisable to ask the parent to summarize his interpretation of the test data so that any misconceptions can be corrected and misunderstandings reduced to a minimum.¹³

Developing Reporting Procedures for Local Use

Since a satisfactory reporting plan needs to be understood and approved by the teachers, students, and parents, it should be developed cooperatively, with all these groups having representation. Any changes in reporting practice should be preceded by a carefully planned program of interpretation to students and their parents so that there will be adequate understanding of the changes and why they are being made. The reporting plan should be consistent with the philosophy of the school. It should be comprehensive, including all the major objectives of the educational program.

The forms and procedures used in reporting to parents constitute merely the channel of communication. The effectiveness of a reporting plan depends largely upon what is communicated—that is, upon the adequacy of the teacher's appraisal of student growth and the skill with which he involves students in the evaluative process.

The techniques that are to be used should be selected in light of the relative importance of the different functions of marking, which vary considerably from the primary grades through college. Certainly in the elementary grades, for example, the informational function is crucial. Both

¹³ Adapted from Gilbert Sax, *The Construction and Analysis of Educational and Psychological Tests: A Laboratory Manual* (Madison, Wisc.: College Printing and Typing Company, 1962), pp. 66–67.

teachers and parents have much information of value to communicate to each other through such a technique as the teacher-parent conference. At the elementary school level, the teacher should certainly not have to place much reliance on the motivating function of marks; and their use in the administrative and guidance functions is usually one of supplementing data from a systematic program of standardized testing in the basic skills. At higher levels, reliance on letter grades becomes increasingly appropriate as the student goes from junior high through senior high and college. The teacher increasingly bases his summary grades on limited samples of academic work; hence, he may have insufficient information to justify the writing of informal letters or holding conferences with all parents. Also the higher the grade level, the greater the importance of the first two functions, which depend on data that is either in numeric form or can readily be translated into such form.

Perhaps no plan of reporting would be quite so effective as periodic teacher-parent and teacher-student conferences. However, such a plan becomes impractical at the high school level unless the number of teacher-student contacts is materially decreased or unless allowance is made for such conferences in the work load of homeroom, core-class, or guidance teachers. In general, Wrinkle's experience has led him to favor the check-list, or rating scale, as a shorthand method of communicating a maximum amount of information to high school students and their parents with a given amount of time and effort and as a means of providing data that can be easily summarized and recorded.¹⁴

One of the essential characteristics of a good reporting plan, especially at the secondary school level, is *flexibility*. That is, the plan should permit core teachers or guidance teachers to report in some detail on the achievement and behavior of students, and yet be so designed that a teacher who has almost two hundred students (for example, in typewriting or physical education) is not obligated to check routinely many aspects of student behavior that he has had no opportunity to observe or record. The growth report used in the Pasadena junior high schools is designed to allow for such variations in practice from teacher to teacher.

Under this plan, each student receives a subject grade and a citizenship grade in each subject. Since both marks are recorded on the cumulative record, they are both considered important by students and teachers. Students realize that many employers will be just as interested in their citizenship record as in their scholarship record. Each of these two grades is defined in terms of several subheadings simply worded so that they are meaningful to students and parents. For example, the subheadings under the citizenship grade are: (1) Responsibility, (2) Effort, (3) Participa-

¹⁴ Wrinkle, *loc cit.*

tion, (4) Class Conduct, and (5) Courtesy. Those teachers who so desire may place a "plus" or "minus" symbol following any subheading to indicate a strength or weakness for that student. Ample space for dated comments by teachers and parents is provided on the back of the report form.

Not only does this report allow for variations from teacher to teacher but it permits a teacher to report much more fully on certain students than on others. Such flexibility encourages the teacher to do as adequate a job of reporting as time permits without forcing him into a pattern of routine checking for large numbers of traits.

IMPROVING THE VALIDITY, RELIABILITY, AND COMPARABILITY OF TEACHERS' MARKS

It seems that letter marks are almost essential to the administrative functions, and that they might serve the guidance functions moderately well if they were more nearly comparable from teacher to teacher. It is important, therefore, that we direct our attention to principles and procedures that might increase the validity, reliability, and comparability of teachers' marks.

As with achievement tests, our major concern with teachers' marks is that they have content validity, that the evidence used in grading be adequately representative of the objectives and content of the course. We are also concerned with the reliability of grades, that is, that the sampling of evidence be large and that subjectivity of judgment be minimized as much as is feasible in assigning scores to the evidence utilized as a basis for grading.

When we consider how we can improve the validity, reliability, and comparability of teachers' final grades, we recognize that three essentially different problems are involved:

1. How to improve the validity and reliability of the *sampling of evidence* used as the basis for grading.
Assistance on this aspect of the problem has been given throughout the textbook as we have considered how to improve the construction and scoring of teacher-made tests, how to use standardized test results as an aid in grading, how to rate products and processes, and the like.
2. How to *weight and combine* the cumulated data for a semester or year into a single composite score for each student, using procedures that (a) reflect the desired emphasis on different types of evidence and (b) are objective, reproducible, and easily explained to students and parents. Grades must not be used to reward those who have especially pleased the teacher and to punish those who have not. The procedures for weighting and combining rollbook data should be public information.
3. How to divide the distribution of composite scores for the assignment of A, B, C, D, and F grades. It is with respect to this aspect of the problem

that we encounter such problems as (a) improving the comparability of grades for multiple sections of the same course; (b) improving the comparability of grades from one subject to another; and (c) improving the comparability of grades from school to school.

Improving the Validity and Reliability of the Sampling of Evidence Used in Grading

If the sampling of evidence cumulated in the teacher's rollbook is to constitute a valid basis for marking, the evidence should represent all the major goals of the instructional program. Moreover, the evidence basic to the subject grade should be limited to information regarding the student's competency with respect to these major goals, uncontaminated by extraneous factors.

Grades recorded on tests or products should be as free as possible of factors extraneous to student achievement. Tests and homework should be scored as objectively as possible; grades on student knowledge in different areas should not be contaminated by such factors as spelling, neatness, or handwriting. Assignments or tests, on which subjective scoring is essential, should be scored by methods that reduce "halo effect." Students should not receive spuriously high marks for reports that have been "dressed up"; clear communication, legibility, and neatness should be sufficient. Nor should students be rewarded for writing reports that are unusually long. The teacher should suggest an acceptable range in length and indicate that writing a longer report will *not* contribute to a higher grade.

In order to improve reliability in marking, a *large* sampling of evidence should be obtained; *objective* scoring should be used whenever it can be used without reducing relevance; the raw scores on short quizzes should be recorded so that evaluative judgments can be based on *combined raw scores*; tests should be of optimum difficulty so as to differentiate among students of different levels of competency.

Weighting and Combining the Data into Composite Scores

If students are to be fairly graded and to be motivated by knowing about their progress to date, the bases for combining data from tests, term papers, and other types of evidence should be clearly defined and reported to the class. Insofar as possible, the student should be able to reproduce this combining process and estimate his grade. It is true that for students near the borderline between one grade and another, the student's score on a final examination, and occasionally subjective judgment, must be the

decisive factor. If the latter is the case, the teacher should be able to tell the student what factors were considered in making that judgment. For example, for students making identical composite scores at the borderline between A and B, the teacher might decide to give an A to the student whose work had improved during the last few months of the year as compared to one whose work had retrogressed.

The use of stanine scores on essays and other products will force the teacher to differentiate among levels of competency and will improve the comparability of marks from one marking occasion to another.¹⁵ Or the teacher can set up any other system of converted scores, provided that it can be explained to students and is consistently followed. Numerical scores are more easily recorded in rollbooks than A, B+, and the like; moreover, they are more easily totalled at time of summarization. Since students like to receive A's or B's, however, the teacher can present his own interpretation of the numerical scores. The author has used the following:

9	A
8	
7	B
6	
5	C
4	
3	D
2	
1	F

The students are instructed to interpret a score of 8 as "between an A and a B"; a score of 6 as "between a B and a C," and the like. Students can easily keep a cumulative record of their own grades if they wish.

If all rollbook data are recorded in terms of comparable scores, and marks for examinations or projects that should receive double weight are recorded twice, all the teacher needs to do at grading time is to total all rollbook entries for each student. Such a total gives the composite score for each student. Or if missing entries due to absences are a problem, one might check off the highest scores for each student until a median or midscore is reached.

If the teacher does not use stanines or some other type of standard score in the rollbook, he cannot easily weight different types of evidence

¹⁵ If the teacher finds it difficult to use the standard percentages in grading all assignments, he could make sure that approximately one-fourth of the students receive high stanine scores (9, 8, or 7); that approximately one-half receive the three middle stanine scores (6, 5, or 4); and that approximately one-fourth receive the low stanine scores (3, 2, and 1).

as desired. The types of scores that have the highest *SD*'s will receive the greatest weight in any composite score.¹⁶

DIVIDING THE DISTRIBUTION OF COMPOSITE SCORES FOR THE ASSIGNMENT OF MARKS Once composite scores have been obtained, the teacher faces the decision of how generous to be in counting off the number of A's, B's, and other marks. It is at this step in the decision process that high school administrators and counselors would like to intervene to improve the comparability of grades. Grades would be more meaningful to all who use them and would have higher predictive validity if greater comparability in grading practices could be achieved.

In Sweden, comparability on a national scale is achieved by administering a common achievement examination to all students in a subject field and then informing each school concerning the number of their students who scored well enough on the test to be entitled to the highest mark and each of the other marks. The centralized examination does not determine *which students* will receive each mark but how many of each mark are available for assignment. In this way, the teachers retain full responsibility for the assignment of marks; they can take into account additional evidence and weight each type of evidence as desired. Centralization of decisions concerning the *number* of A's, B's, and the like to be assigned ensures that grades have high comparability from school to school.

Centralized control of marking, however, may have disadvantages. The common test, unless it is very carefully constructed, may not fairly represent the major goals of instruction; some schools might be allowed to give more high grades because their curricular emphases approximated closely those of the common examination. Negative effects on students' motivation and their self-concepts in schools with larger percentages of slow-learning pupils seem inevitable. Since American schools and colleges use standardized tests as aids in admission and classification, there is less need in the United States than in many other countries to ensure comparability of grades from school to school.

In an effort to increase comparability of grades *within* the school, many

¹⁶ For example, a teacher may tell students that the greatest weight will be given to homework in determining the final grade. However, if there is a relatively little variation in grades assigned to homework and much greater variation in grades on periodic examinations, the examinations will automatically contribute more to the final mark if any objective method of combining data is used. The *effective weight* of any component increases with its variability. If stanine scores or *T*-scores are used for all variables to be combined, this problem can be ignored for the variability for all components in the grading composite will be the same. If comparable scores are not used, the desired weight for each type of evidence (homework, term paper, examinations, and the like) must be divided by the *SD* of scores for that type of evidence before the data are combined to obtain standard scores.

schools and colleges establish school-wide grading policies. If we wished to help teachers to modify the school grading policy for various sections of English or some other multiple-section course, an "anchor test," such as that described in Chapter 13, might be developed. Teachers might then be informed concerning how each of their classes performed on this anchor test.¹⁷ The distribution of students' scores on the anchor test, however, would not *determine* the number of A's and other grades available but would merely help teachers to modify the school grading policy for specific classes on an objective basis.

If teachers did not wish to use an anchor test, the school grading policy might be adapted to specific classes in terms of how well the students achieved on some test that was highly correlated with achievement in the subject field. In measurement terms, we would select a test that would have high concurrent validity, that is, high correlation with the criterion of over-all achievement in the course. Scannell¹⁸ has recommended the use of such a test as the *Cooperative English Test* in English classes and other relevant tests for other departments.

Let us consider how Scannell's proposal might work. Let us assume that all ninth-grade students have taken a local arithmetic test and a scholastic aptitude test with verbal and numerical subtests. Let us assume that *T*-scores on the arithmetic test and the numerical section of the scholastic aptitude tests have been used by counselors to aid them in programming eighth-grade students into basic mathematics (a remedial course) general mathematics, and algebra. The distributions of student's average *T*-scores in these two tests are shown in Table 16.1. Let us assume that the grading policy for the school is to assign approximately 20 percent A's, 30 percent B's, 40 percent C's, and 10 percent D's and F's. We have computed P_{80} , P_{60} , and P_{40} for the algebra classes, the nonalgebra classes, and all classes combined since these are the division points that divide each group into the recommended percentages, that is, highest 20 percent, next highest 30 percent, and the like.

Routine application of this grading policy in all classes is indefensible because of differences in their achievement. Such questions as the following cannot be avoided: Is a specific class representative of the general school population? If a student is assigned to an honors section, should he be penalized by the fact that his competency is below average in that class, although it would be well above average in a typical class in the

¹⁷ If the teacher were using such an anchor test as part of his final examination, he could supplement the anchor test with a number of difficult items of his own selection for accelerated classes. For a slow-learning group the anchor test could be supplemented by a number of relatively easy items.

¹⁸ Dale P. Scannell, "Making Grades Meaningful: A Proposal," *The University of Kansas Bulletin of Education*, vol. 15 (November 1960), pp. 26-35.

Table 16.1
Frequency Distribution of Ninth-grade Mathematics Students with
Respect to Scores Used as a Partial Basis for Grouping

AVERAGE <i>T</i> -SCALED SCORES ON ARITHMETIC TEST AND NUMERICAL SECTION OF SCHOLASTIC APTITUDE TEST	Number of Cases in						TOTAL 9TH GRADE CLASS
	BASIC MATH.	GENERAL MATH.	TOTAL NON- ALGEBRA	REGULAR ALGEBRA	HONORS ALGEBRA	TOTAL ALGEBRA	
70 and above				1	9	10	10
69					2	2	2
68				1	2	3	3
67					5	5	5
66		1	1	2	5	7	8
65				3	4	7	7
64		1	1	2	3	5	6
63		2	2	4	4	8	10
62	1	1	2	3	2	5	7
61		3	3		4	4	7
60	1	3	4	2	4	6	10
59		3	3	6	3	9	12
58		4	4	8	2	10	14
57	1	2	3	10	1	11	14
56		1	1	16		16	17
55		2	2	16		16	18
54		1	1	16		16	17
53	1	3	4	15		15	19
52		4	4	15		15	19
51	2	9	11	7		7	18
50	1	12	13	7		7	20
49	2	15	17	5		5	22
48	4	13	17	5		5	22
47	3	15	18	2		2	20
46	4	12	16	2		2	18
45	5	13	18				18
44	6	10	16	1		1	17
43	7	10	17	1		1	18
42	8	7	15				15
41	7	6	13				13
40	5	7	12				12
39	6	5	11				11
38	5	5	10				10
37	6	3	9				9
36	5	4	9				9
35	6	2	8				8
34	5	3	8				8
33	4	2	6				6

AVERAGE <i>T</i> -SCALED SCORES ON ARITHMETIC TEST AND NUMERICAL SECTION OF SCHOLASTIC APTITUDE TEST	Number of Cases in						TOTAL 9TH GRADE CLASS
	BASIC MATH.	GENERAL MATH.	TOTAL NON- ALGEBRA	REGULAR ALGEBRA	HONORS ALGEBRA	TOTAL ALGEBRA	
32	2	2	4				4
31	3	1	4				4
30	1	1	2				2
Below 30	9	2	11				11
Total number of cases	110	190	300	150	50	200	500
P_{80}			49			63	58
P_{50}			44			56	49
P_{10}			34			50	36
<i>Ninth grade as reference group</i>							
A's (Percent at P_{80} and above)	2	9		21	98		20
B's (Percent between P_{50} and P_{80})	6	26		71	2		30
C's (Percent between P_{10} and P_{50})	65	58		7			40
D's and F's (Percent below P_{10})	27	7					10
<i>Nonalgebra and algebra classes as reference groups</i>							
A's (Percent at P_{80} and above)	8	35		9	68		
B's (Percent between P_{50} and P_{80})	20	33		30	32		
C's (Percent between P_{10} and P_{50})	55	27		51			
D's and F's (Percent below P_{10})	17	4		11			

same subject? Should larger percentages of A's be assigned in honors classes and in other classes, such as physics or solid geometry elected by able, college-bound students? Certainly, each teacher's following a school grading policy in each of his classes without making allowance for differences in competency among groups is unjustifiable. The question, in

measurement terms, becomes, What reference group should be used as the basis for defining the A-level and other levels of achievement?

It would seem that the number of students in a class that had average T -scores above 58 (P_{80} for all ninth-graders) on the tests used in grouping might be the approximate number to receive A's. If this basis were used, the number of A's could vary considerably from class to class, but there would still be about 20 percent A's in "all classes combined." Let us assume that Mr. Brown has one class in basic mathematics, another in general mathematics, one regular algebra class, and an honors algebra class. Let us further assume that each of these classes is representative of all students enrolled in these subjects. Then if we used the entire ninth grade as our reference group, the percentage of each mark would be approximately as follows:

	Percentage of Students Assigned Each Mark in Mr. Brown's Classes			
	BASIC MATHEMATICS	GENERAL MATHEMATICS	REGULAR ALGEBRA CLASS	HONORS ALGEBRA CLASS
Percent of A's (P_{80} and above in related tests)	2	9	21	98
Percent of B's (P_{50} - P_{80})	6	26	71	2
Percent of C's (P_{10} - P_{50})	65	58	7	
Percent of D's and F's (below P_{10})	27	7		

At least two problems arise when a procedure like this is followed:

1. The algebra students may be graded too liberally in terms of the population of students with which they will be competing in college and adulthood; for example, in this hypothetical situation, 93 percent of the regular algebra students would receive "recommended grades," which communicate to students, parents, and college authorities evidence of relative strength in mathematics.
2. Students who spend most of their academic lives in groups that consist of average and below-average students are almost condemned to receiving low marks, regardless of the amount of effort they expend or the amount of growth they make toward the objectives of the course.

In the light of this illustration, it seems undesirable that the reference group be the general school population. It may be more desirable for the reference group to be all students enrolled in a *subject field*. If grades are reasonably comparable *within* a subject field, they can serve adequately for the first, second, and fourth purposes (as noted on page 512) without having negative effects upon the motivational level or self-concepts of students of average or below-average scholastic aptitude. In other words,

we might use multiple reference groups and rely upon the fact that those who need to interpret marks, such as admissions officers and counselors, know that basic mathematics is usually concerned with remedial work in arithmetic and that general mathematics students tend to be a less select group than algebra students.

In other words, differentiated courses could be set up and clearly identified by name, for example, the remedial mathematics course would be called basic mathematics, while accelerated mathematics courses would be labeled on the cumulative record as honors courses. Then, if a school defined an A as representing the top 20 percent of its students with respect to competency in any subject field, approximately 20 percent of the students in basic, regular, or honors mathematics who excelled in their progress toward the objectives of the course could receive A's. Under such a plan, however, a conscientious plodder in remedial courses might conceivably become valedictorian; while students would shun honors courses because of the difficulty of attaining the position of top 20 percent in such courses.

It seems that the routine application of either of these plans (that is, using the total school, or enrollees in a specific subject, as the reference group) leads to undesirable complications. Since teachers vary widely in their generosity, however, some sensible compromise needs to be reached so that grades can have greater comparability than they do at the present time. A reasonable compromise might be to use Scannell's general approach in modifying the general grade distribution for specific classes *except* that the reference group for college-preparatory classes should be the college-bound students (with whom they will be competing) while the reference group for classes that are not college preparatory should be the remainder of the student body. In other words, the number of students in a specific class scoring above P_{80} for the nonalgebra classes would determine the approximate number of A's in basic mathematics and general mathematics; while the combined algebra classes would constitute the reference groups for both regular and honors sections of algebra. If this procedure were followed, the percentages for Mr. Brown's classes might be as follows:

Percentage of Students Assigned Each
Mark in Mr. Brown's Classes

	BASIC MATHE- MATICS	GENERAL MATHE- MATICS	REGULAR ALGEBRA CLASS	HONORS ALGEBRA CLASS
Percent of A's	8	35	9	68
Percent of B's	20	33	30	32
Percent of C's	55	27	51	
Percent of D's and F's	17	4	11	

According to this modification, an honors or regular class of college-bound students would be compared with a reference group of college-bound students. Students would not shun honors courses because of the probable negative effect on their grades. If a specific class in algebra had a higher proportion of able students than other classes, a larger-than-average percentage of A and B grades could be assigned. Teachers would have complete freedom in deciding to which students each mark should be assigned. In fact, modification in the standard distribution of grades would be recommended rather than imposed. Certainly, the number of students who receive failing grades should not be determined by formula; nor should all students in honors classes receive A or B grades if their work does not justify such marks. An attempt would be made to convince the teachers of the advantage of voluntary cooperation in improving the comparability of grades.

The proposals made in this chapter section have been based on the premises that (1) single classes cannot logically be graded on any school-wide grading policy and (2) grading policies should be modified for specific classes on the basis of interclass differences with respect to some test (or tests) maximally related to differences in student competency in the subject. If the adjustment can be made on the basis of a common achievement test or anchor test, we have the advantage of using a variable maximally related to achievement in the course. Otherwise, we can make the adjustment on the basis of some variable that has high concurrent validity as a substitute for such an anchor test.

The use of a related test or tests has at least two advantages over the use of an anchor test: (1) teachers are not threatened by the possibility that class differences in average test score might be interpreted as representing differences in teaching effectiveness and (2) information concerning the approximate distribution of marks would be available at the *beginning* of the year so that grades assigned during the year could be consistent with final grades. If, in terms of the best data available, one class seems likely to merit approximately 10 percent A's and another 30 percent A's, differences in grading policies could be initiated early. These data from related tests, however, should constitute only an aid to the teacher's professional judgment; whenever the teacher can devise more valid bases for modifying school grading policies, he should be encouraged to use them and to share them with his coworkers. One disadvantage of using related tests (as a basis for modifications of school grading policy) is that the plan fails to take into account differences in teaching effectiveness in facilitating student learning during the semester or year; whereas such differences would be reflected in a final examination or anchor test.

Perhaps the best way for an administrator to help teachers to adapt standard grading policy to their own classes would be as follows: (1)

translate the grading policy regarding percentages of each mark into percentile scores that represent these percentage ranges; (2) translate these percentile scores into equivalent scores on anchor tests or related tests for college-bound and other students; (3) prepare a report showing the percentage of cases in each class falling within those limits. These would constitute recommended percentages of A's, B's and the like for each class. For example, Mr. Brown would receive a sheet for algebra classes in which his honors and regular classes would be shown along with all college-preparatory classes in mathematics. He would receive another sheet in which his classes in basic and general mathematics would be shown in comparison with all noncollege-preparatory classes in mathematics.

It is recognized that a number of high school courses cannot readily be classifiable as college-preparatory or not. All students, for example, are required to take American history. If classes in such subjects are sectioned, the higher-ability sections could be classified as college bound. In subjects like typewriting, art, music, home economics, and industrial arts, the teacher's subjective judgment about whether a class was a relatively high-achieving or low-achieving class might constitute the best basis for modifying the general grade distribution. Over a period of several semesters, however, a teacher's deviations from school grading policy should average out so that his cumulated distribution of marks would approximate the recommended distribution.

SUMMARY STATEMENT

A functional cumulative-record system is essential if measurement and evaluation data are to make their maximum contribution to the understanding of students and their problems. Cumulative records can be used to provide a longitudinal picture of each student's development, to help teachers in understanding each student's special needs and problems, and to assist school personnel in helping the parent to achieve a more objective picture of his son or daughter. In addition, they constitute the official record of a student's attendance, scholarship, promotions, and graduation.

A good cumulative-record system should have the following characteristics: (1) A cumulative record should be maintained for each student. (2) The record should be transferred with the student (at least in summary form). (3) The record should be comprehensive. (4) The forms should be simple and easily understood. (5) The cumulative-record system should be flexible, requiring a minimum of data for all students but permitting great latitude in the types of additional data cumulated. (6) The records should be designed to reveal trends in growth over a period of years. (7) Although the records should be readily accessible to teachers, the confidential nature of the data must be respected. (8) Cumulative records should be so maintained that the data are accurate, complete, and up-to-date. (9) In recording data, every attempt should be made to distinguish facts from personal opinions.

In order for the data in cumulative-record folders to be most valuable, they should be organized for ready accessibility and culled out periodically. A summary sheet is useful for indexing materials in the cumulative-record folder.

Programs for reporting grades and other evaluation data should be appraised in terms of how well they serve administrative, guidance, motivational, and informational functions. Perhaps the most important function of any reporting plan is to provide the information necessary for sound, cooperative planning by students, parents, and teachers.

New developments in reporting practices include: (1) changing the symbols; and (2) supplementing the symbols by (a) the development of fairly detailed rating scales for major objectives on which the teacher rates the student, (b) informal letters, and (c) parent-teacher conferences.

In improving the validity, reliability, and comparability of teachers' marks, three essentially different problems are involved: (1) how to improve the validity and reliability of the sampling of evidence used as the basis for grading, (2) how to weight and combine the cumulated data for a semester or year into a single composite score for each student, and (3) how to divide the distribution of composite scores for the assignment of A, B, C, D, and F grades.

SELECTED REFERENCES

- ALEXANDER, WILLIAM M., "Reporting to Parents—Why? What? How?" *National Education Association Journal*, vol. 48 (December 1959), pp. 15–28.
- CAGLE, DAN F., AND RAY C. HEISCHMAN, "How May We Make the Evaluation and Reporting of Student Achievement More Meaningful?", *Bulletin of the National Association of Secondary School Principals*, vol. 39 (April 1955), pp. 24–30.
- DOBBIN, JOHN E., "What Parents Need to Know About Tests and Testing," *National Elementary Principal*, vol. 34 (September 1954), pp. 152–160.
- DUROST, WALTER N., "How to Tell Parents about Standardized Test Results," *Test Service Notebook*, No. 26. New York: Harcourt, Brace & World, Inc., 1961. Available on request.
- HARRIS, F. E., "Three Persistent Educational Problems: Grading, Promoting, and Reporting to Parents," *Understanding the Child*, vol. 23 (April 1954), pp. 34–42.
- HORST, PAUL, "How Much Information on Test Results Should Be Given to Students: Views of a Research Psychologist," *Journal of Counseling Psychology*, vol. 6 (Fall 1959), pp. 218–222.
- KELLY, ELDON C., "A Study of Consistent Discrepancies between Instructor Grades and Term-End Examination Grades," *Journal of Educational Psychology*, vol. 49 (December 1958), pp. 328–334.
- LACEY, OLIVER L., "How Fair Are Your Grades?", *American Association of University Professors Bulletin*, vol. 46 (September 1960), pp. 281–283.
- LAFRANCHI, EDWARD H., "High School Marks: Comparative or Individual," *School Executive*, vol. 71 (July 1952), pp. 51–54.
- PLOGHOFT, MILTON, "The Parent-Teacher Conference as a Report of Pupil Progress: An Overview," *Educational Administration and Supervision*, vol. 44 (March 1958), pp. 101–105.

- ROTHNEY, JOHN W. M., *Evaluating and Reporting Pupil Progress, What Research Says to the Teacher*, No. 7. Washington, D.C.: National Educational Association, 1955.
- SCANNELL, DALE P., "Making Grades Meaningful: A Proposal," *The University of Kansas Bulletin of Education*, vol. 15, No. 1 (November 1960), pp. 26-35.
- WAHLQUIST, G. L., "How Machine Processes Save Counselor Time," *California Journal of Secondary Education*, vol. 32 (November 1957), pp. 442-445.
- WALTON, WESLEY W., "The Electronic Age Comes to the Schoolhouse," *Systems for Educators*, vol. 8 (January-February 1962), pp. 3-6.
- WECKLER, NORA, "Problems in Organizing Parent-Teacher Conferences," *California Journal of Elementary Education*, vol. 24 (November 1955), pp. 117-126.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. What information about his students should a teacher obtain from their cumulative records at or before the opening of school?
2. Obtain a cumulative record form designed for use at the secondary-school level. Summarize and classify the kinds of information requested.
3. Evaluate the cumulative-record system of a school district in terms of the characteristics listed in this chapter.
4. Draft a one-page summary sheet on which a teacher could summarize the data filed in one pupil's cumulative record folder over a three-year period.
5. To what extent should students' test results be reported to parents? What guidelines might be developed to minimize the problems involved in such a reporting process?
6. Study the bulletins and other materials developed by a school district to assist teachers in conducting teacher-parent conferences (for the purpose of reporting on student growth and planning cooperatively for continued improvement). Summarize the major principles emphasized.
7. In your subject field, what types of data should the teacher have in order to appraise a student's "total achievement" as a basis for assigning marks?

Using Measurement Data in Individual and Group Guidance

Guidance usually includes such basic functions as (1) helping the student to ascertain, understand, accept, and apply the relevant facts about himself in relation to facts about educational and vocational opportunities;¹ (2) maximizing his adjustment to his educational opportunities in terms of his abilities, interests, and needs;² and (3) helping him to reach workable solutions to a variety of adjustment problems.³

It is evident from a consideration of these functions that guidance involves: (1) work with students in every aspect of the school program, (2) knowledge concerning the individual student's abilities, interests, needs, and adjustment problems; (3) the use of a wide variety of formal and informal techniques for understanding students and helping them to understand themselves; (4) obtaining, recording, organizing, and interpreting many types of test and nontest data; (5) skill in helping students to define their problems and to use more effective problem-solving techniques in working out their own constructive solutions to them; and (6) the cooperation of parents and the teamwork of every member of the school staff, together with the staff resources of psychological clinics, character-building organizations, family social-work agencies, law-enforcement groups, and the like.

A concern for obtaining and using data for guidance purposes has

¹ Donald E. Super and John O. Crites, *Appraising Vocational Fitness by Means of Psychological Tests* (New York: Harper & Row, Publishers, Inc., 1962), p. 2.

² C. C. Ross and Julian C. Stanley, *Measurement in Today's Schools* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1954), p. 370.

³ John G. Darley and others, "The Functions of Measurement in Counseling," *Educational Measurement* (Washington, D.C.: American Council on Education, 1951), p. 68.

permeated this textbook, especially Chapter 1, in which some of the problems facing counselors were considered, all the chapters of Part Two on "The Study of Individuals," and Chapter 14, on "Educational Diagnosis."

No consideration will be given in this chapter to such guidance techniques as the counseling interview or the dissemination of occupational information. There will be an attempt only to provide an overview of guidance responsibilities, an analysis of the functions usually performed by teachers and by specialized guidance workers, and a discussion of the issues and principles involved in the use of evaluation data in individual and group guidance, especially as both approaches are concerned with problems of educational planning and vocational choice.

GUIDANCE RESPONSIBILITIES OF COUNSELORS AND TEACHERS

Relationships of Guidance and Education

The concept of guidance is often interpreted so broadly as to become almost synonymous with good education. Such interpretations are basically sound in their emphasis on the teacher as a key guidance worker; in their claim that many guidance problems can be met more constructively by curricular change and individualization of instruction than by counseling students on problems created by the lack of such measures; and in their recognition that guidance presupposes knowledge of the individual and that no full-time counselor can know adequately the needs of the 300 to 800 students who may be assigned to him.

The concept of guidance as synonymous with good education is misleading, however, when used as a basis for implying that *all* guidance can be done by classroom teachers, ignoring the value of specialized training for certain guidance responsibilities, or minimizing the need for coordination of the school guidance program. The increasing complexity of modern society has made it necessary for students to have more assistance in making wise choices among diverse opportunities in the occupational world and among the varied curricular, extracurricular, and work-experience opportunities in high school and college programs. As one guidance textbook succinctly states it, the concept of guidance as synonymous with good education "will probably contribute far more to good teaching than to improved guidance."⁴

⁴ D. Welty Lefever, Archie M. Turrell, and Henry L. Weitzel, *Principles and Techniques of Guidance* (New York: The Ronald Press Company, 1950), p. 23.

The Roles of Specially Trained Personnel and of Teachers in the Guidance Program

The best division of responsibilities for guidance must vary with the size of the school, the specialized staff, and facilities of the school system and the community in which it is located, the attitudes and training of the administrative-staff members, and many other factors. Many textbooks in guidance include a variety of organizational charts for small high schools, large high schools, and those high schools located in large urban communities with a wealth of specialized resources available.

Klausmeier has approached this issue by attempting to distinguish between (1) six major services required in an organized guidance program and ordinarily performed by specially trained counselors or guidance coordinators, and (2) services ordinarily performed by teachers.

SERVICES USUALLY PERFORMED BY SPECIALLY TRAINED PERSONNEL
The six major services usually performed by specially trained personnel are as follows:

1. *Coordination and leadership in the appraisal program*, assisting teachers in securing and using evaluation data to understand students as individuals.
2. Providing leadership in a program of *collecting and disseminating accurate occupational information* to students.
3. *Providing counseling services* (including direct counseling to students and in-service education of teacher-counselors).
4. *Coordinating group approaches* to guidance.
5. *Coordinating the referral program* and maintaining effective working relationships with out-of-school agencies.
6. *Directing research* directly concerned with the guidance program⁵ [Italics added.]

Klausmeier emphasizes that these functions may be performed in the small schools by the principal and counselor, and the large schools by the guidance coordinator, psychologist, psychometrist, and a staff of counselors.

THE TEACHER AND GUIDANCE SERVICES As teachers increasingly develop a guidance point of view and a well-rounded background for guidance responsibilities, administrators often find it desirable to spread guidance functions among a number of teachers, thus reducing the number

⁵ Adapted from Herbert J. Klausmeier, *Principles and Practices of Secondary School Teaching* (New York: Harper & Row, Publishers, Inc., 1953), p. 411.

of student contacts for each adviser and facilitating the development of a more personal adviser-student relationship. Many teachers—homeroom and core-class teachers as well as teachers of occupations and other special guidance classes—have been assigned specific responsibilities for group approaches to guidance or for individual counseling. Almost all teachers have some responsibilities that involve significant potentialities for guidance (for example, sponsoring an art club; advising the student council, newspaper, or annual; or teaching remedial reading).

The guidance aspects of *regular* classroom instruction are becoming increasingly significant. One should not underestimate the importance of the total contribution made to guidance by those classroom teachers who know their students individually and attempt to meet their personal, as well as academic, needs. The guidance aspects of regular classroom instruction may be summarized as follows:

1. *Appraising individual students*

There should be a two-way exchange of information between counselors and teachers:

- a. Teachers being supplied with pertinent data from the cumulative record, and
- b. Teachers feeding into counselors and the cumulative-record system such materials as anecdotal records, reports of interviews, and periodical summaries of qualitative judgments that aid in building a longitudinal picture of the student's growth and his changing needs.

2. *Helping students to discover their aptitudes and interests*

As an integral part of a good instructional and cocurricular program, the teacher helps each student in discovering his own strengths and weaknesses, both with respect to his past achievement and with respect to any special aptitudes that may have significance for educational and vocational planning.

3. *Practicing informal personal guidance in the classroom*

Informal guidance contacts include discussing with students probable reasons for poor scholarship and necessary steps toward improvement, and holding many informal interviews with students when they come in with personal problems after school, or in club activities that the teacher sponsors, and the like.

4. *Creating success experiences and helping students to interpret and build constructively on necessary failure experiences*

- a. Through adaptation of school tasks to individual strengths and weaknesses, and through the provision of prestige-giving experiences for students
- b. Through the use of sociometric techniques to place students in group situations in which they have improved chances for succeeding socially
- c. By helping students to interpret necessary failure experiences as a basis for future growth and to prevent or minimize future failure experiences.

5. *Disseminating occupational information relevant to the subject taught*
6. *Identifying students who need special help and referring such students to specialized guidance personnel*

Since teachers are the only staff members who have regular daily contacts with students (with opportunities to observe them in peer relationships, experiences of success and failure, attitudes toward assigned tasks, and other relationships which reveal adjustment problems), they have a special responsibility for identifying and referring those students who need assistance from specialized guidance personnel.

This analysis of guidance functions is intended to help clarify the inevitable overlapping of, and close interrelationships between, the guidance functions of teachers and of specialists, as well as the need for coordination and leadership.

ISSUES AND PRINCIPLES INVOLVED IN THE USE OF MEASUREMENT DATA IN GUIDANCE

The foregoing review of the guidance responsibilities of school personnel has undoubtedly called to mind many techniques of measurement and evaluation that can aid students and counselors in their cooperative attack upon guidance problems. Knowledge of measurement data on students' cumulative record forms and familiarity with the contents of their record folders would appear to be indispensable in many aspects of the guidance process. It may seem strange, therefore, that the function of measurement in guidance and the role of the counselor in presenting data to high school students are controversial issues in guidance.

Issues Concerning the Advisability of Interpreting Test Data to Students

Those who question the advisability of interpreting test data to students tend to base their position on several premises, each of which will be considered.

1. *That the influence of unmeasured factors (notably the student's "drive" or motivation) is so great that the measurable characteristics, by comparison, are relatively insignificant.* As tests and aids to their interpretation have improved and as counselors have learned to consider significant nontest data in combination with test results, this argument is less frequently offered. The problem is essentially one of deciding which aspects of behavior can be measured by tests, selecting the best tests for the purpose, and appraising, as objectively as possible, the other factors.

The significant factor of "drive" or motivation must be kept constantly

in mind. Although the counselor has no score on the student's motivation, he does have useful *data*—reflected in student marks, teacher comments, participation in curricular and work-experience activities, and the like.

Let us imagine, for example, that a student who is seeking the counselor's advice on college attendance has shown good motivation and effort throughout his school years (as reflected in teachers' comments in his cumulative record folder and a consistent record of B and C grades). His IQ's on three group tests range from 88 to 100. Even with unusually good effort, this student has done only average work in high school. Such data probably constitute an adequate basis for the counselor's encouraging this student to consider several vocations that require no college preparation. Interest inventories, aptitude-test results, records of marks and activities, and the student's own expressed interests in the interview situation would provide leads which the counselor could use to approach the problem positively, rather than negatively.

Let us imagine another student, with a consistent record of IQ's above 120 but a spotty record of grades, showing relatively poor motivation. If this student has good marks in his field of special interest, and appears *now* to be highly motivated by his assignment as a laboratory assistant in science, the counselor probably has sufficient data for encouraging him to attempt college work in science, especially if the economic status of his family will permit him to devote full time to his studies.

In both these cases, data on untested factors of motivation and economic status should be considered along with test data.

2. *That existing tests have serious limitations.* This criticism is a valid and exceedingly important one. Persons who reject *all* test data on this basis, however, are probably reacting strongly against the tendency of students, and even of some counselors, to place too great faith in test data, that is, to feel that they will automatically provide the answers to problems. Test data must always be used as an aid to professional judgment, not as a substitute for it.

Super and Crites emphasize the limitations of available tests and the necessity for interpreting test data cautiously, in the light of all other available information.

The use of tests by a vocational counselor is . . . of necessity generally not a predictive process but rather a clinical procedure. A variety of data have to be studied in relation to each other, and hypotheses are established. . . . It should be noted that the term *hypotheses* is used, rather than *conclusions*, as their bases are not definite enough to warrant the term *conclusion*.⁶ [Italics added.]

⁶ Super and Crites, *op. cit.*, p. 533.

Those who "view with alarm" the limitations of test data often fail to consider the more serious limitations of working on the basis of subjective judgments that may have little or no basis in fact. Four valuable properties of test data, as compared with more subjective judgments, can be cited:

- a. The property of *reliability, accuracy, and objectivity* (which serves as a necessary antidote to tendencies toward over- or underestimation of qualifications by either counselor or student).
- b. The property of *validity or meaning or predictive power*. Good psychological tests, despite their limitations, have predictive value for vocational and educational success. This predictive value can serve as a balance wheel in the wishful thinking with which students approach many decisions.
- c. The property of *economy of effort*. As short, standardized samples of behavior, tests can often supply in a short time and at a relatively low cost a basis of judgment-making that is a practical substitute for trial-and-error decisions.
- d. The final property is their *normative* aspect, or their indication of an individual's standing relative to others of similar age, background, or experience.⁷

3. *That the use of test data in guidance tends to make the student a passive, dependent receiver of information, rather than an active solver of problems.* Test data can be presented in such a way as to create a passive mind-set on the part of the student. If the student expects the test data to provide specific answers to his problems, the counselor must stress the limitations of tests and the way in which test and nontest data should be checked against each other in making judgments. If test results are used appropriately in the counseling process, that is, merely to provide needed information and to question or to verify hypotheses, student passivity and dependence will be avoided. The student should be encouraged to interpret his own results, checking his interpretations with the counselor's judgment, and evolve his own hypotheses concerning their implications for his plans.

4. *That interpretation of test results by the counselor may threaten the student's concept of self and hence disturb the counseling relationship.* This criticism represents a very real hazard and emphasizes the need for great skill and caution on the part of the counselor. Growth in student self-understanding, however, is a very important goal of the guidance program and one that should be attacked skillfully, rather than evaded. In a significant study, test data were interpreted to high school students in individual conferences by highly qualified counselors under good conditions of counselor-student rapport. Rothney reported that 60 percent of the student reactions to the test data, presented under these favorable

⁷ Adapted from Darley and others, *op. cit.*, p. 77.

conditions, were definitely favorable ("expectations or current plans confirmed," "results higher than expected," "seemed pleased," "showed high interest"); 9 percent were clearly negative ("disappointed," "skeptical"). The remaining 31 percent of the reactions, although they might be classifiable as neutral, implied indirect rejection.⁸ The process of gaining insight is *emotional* rather than *rational*.⁹ The student's reactions to objective evidence need to be sounded out and examined in a manner characteristic of the nondirective interview.

In interpreting test results and in presenting or summarizing other information, the counselor employs an objective, impersonal approach, and seeks to increase the client's self-acceptance. . . . [When information is presented] in an impersonal, objective manner, the individual must unconsciously go through a process of giving it personal meaning, and it is this "process of giving meaning" which the counselor may use to help his client know himself better. Reported case-histories show that when this impersonal presentation has been used, significant self-analysis occurs.

No matter how perfectly standardized a test may be, the test results will be of no more use than the client can allow them to be. Assimilating and making use of test information is a problem of feelings and attitudes. . . . Letting him proceed at his own pace, stating his objections or approval, examining why he objects or approves, expressing his feelings freely, will prove more valuable in the long run.¹⁰

The student may either be able to accept the test predictions and use them in his thinking or he may need to distort them in some degree. The more the student feels free to discuss his reactions to test results with his counselor, the more likely it is that he will be able to comprehend them and accept their probable significance for him.

Principles Involved in Interpreting Test Data to Students

Much of the distrust concerning the use of test data in guidance has developed from questionable practices in this area. As a basis for approaching the problem more positively, the authors have summarized a number of principles for the use and interpretation of measurement data in guidance.

⁸ John A. M. Rothney, "Interpreting Test Scores to Counselees," *Occupations*, vol. 30 (February 1952), pp. 320-322.

⁹ B. J. Covner, "Nondirective Interviewing Techniques in Vocational Counseling," *Journal of Consulting Psychology*, vol. 11 (March-April 1947), pp. 70-73.

¹⁰ Richard W. Kilby, "Some Vocational Counseling Methods," *Educational and Psychological Measurement*, vol. 9 (Summer 1949), pp. 186-188.

1. The best available tests for the purpose should be used. Selection should be based on the major criteria of validity, reliability, and adequacy of norms, rather than on such minor criteria as cost or ease of scoring. In the selection of aptitude tests, the availability of suitable norms and meaningful validation studies is especially important. If achievement tests are used as predictors, they should be appraised in terms of their value for this purpose.
2. Responsibility for interpretation of test data, and the in-service education of teachers in test interpretation, should be placed in the hands of specially trained guidance workers.
3. Test data should be considered in the context of all other available information. Although one can interpret extreme test scores with considerable confidence, if they are consistent with other data, one must actively seek new information for students for whom predictor test scores seem to indicate about a 50-50 chance of gaining a desired objective.
4. Test data should be interpreted in terms of probabilities, rather than certainties. Instead of saying, "Jim's aptitude-test score is so low that he will fail in algebra," one should say, "Seven out of ten students with aptitude-test scores as low as Jim's have made D or F grades in algebra." Expectancy tables (such as Table 17.1 and Figure 17.1) involving one or more predictor variables, can be very helpful in this respect.
5. Counselors should present test data to students in such a way that (a) the data are brought into the counseling interview only as they help in meeting a need or attacking a problem formulated by the student, (b) they are presented objectively and impersonally by the counselor with the student

Table 17.1
Expectancy Chart (Based on Grades Received by 500 Students Taking
Arithmetic Test at End of Eighth Grade)

RAW SCORES ON ARITHMETIC TEST	Chances in 100 of Making an Algebra Grade of			
	D OR F	C	B	A
90-99		2	23	75
80-89	5	15	40	40
70-79	10	30	40	20
60-69	30	40	30	
Below 60	80	20		

Note: The grades received by students at each raw-score level (for example, 90-99) are tallied; the frequencies for each row are then translated into percentages so that the predictive validity of the test can be interpreted in terms of chances in 100 of making each grade. Note that since grades represent only an ordinal scale (see Chapter 2, page 63), a correlation coefficient could not be computed by the Pearson product-moment method. Ordinarily, the contingency coefficient would be computed if a validity coefficient were desired. See Quinn McNemar, *Psychological Statistics*, third ed. (New York: John Wiley and Sons, Inc., 1962), pp. 198-202.

interpreting their personal meaning for him, and (c) the student is encouraged to take an active role, expressing his reactions to the test results, relating them to relevant information about his in-school and out-of-school experiences, and formulating hypotheses about their implications for his choices.

6. Guidance workers should use special caution in interpreting data from all tests on which the examinee can falsify or distort his responses at will—for example, tests of interests, personal adjustment, and the like.
7. The counselor's approach in an interview involving test interpretation should be conditioned by his realization that the student's interpretation of test data is sometimes more of an emotional than a rational process. It should be recognized that modifying a student's self-concept so that it becomes more realistic is a gradual process. Hence, his realization of the *implications* of test data should be the outcome of several counselor-student contacts, rather than a single dramatic event. Between these successive contacts, the student can be engaged in reading and other exploratory activities in vocational guidance and other relevant areas.

GUIDANCE IN EDUCATIONAL AND VOCATIONAL PLANNING

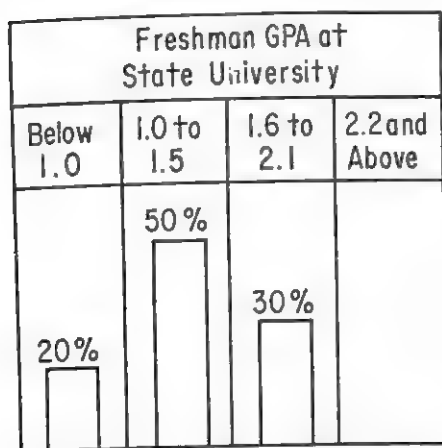
Special attention will now be given to the use of measurement data in connection with Problems 6 and 7 of Chapter 1, that is, helping students with their decisions in educational and vocational planning. A number of the procedures presented are also usable in employee classification.

Counseling as a Cooperative Problem-Solving Process

A central aspect of guidance, and one in which measurement can be of great assistance, is that of counseling students concerning "next steps" in educational and vocational planning. A conference between a counselor and counselee on next steps in life planning should represent a *cooperative problem-solving process*. The counselor can bring much to this conference in terms of his maturity, his experience in aiding in the decision-making process, his resources (in terms of data about the student and about environmental factors that should be considered), as well as his special skills in the interpretation of such data. The student, however, has data about himself (his aspirations, and needs and feelings) that cannot be measured and filed); moreover, he has the basic responsibility for making decisions about "next steps" in his life planning. He is the one who will enthusiastically or half-heartedly test out the hypotheses made about "next steps," or perhaps reject them altogether. He is the one who will live with the consequences of his decisions.

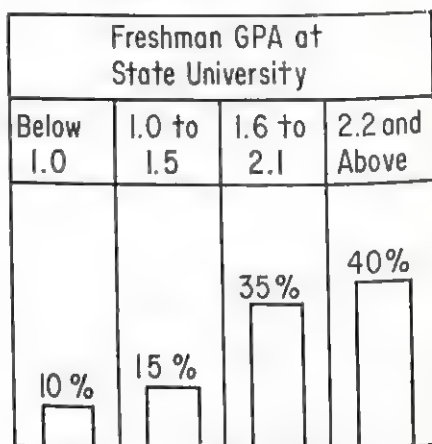
Graphs for Students Scoring in Stanines 8 and 9 (IQ's 120 and above)^a

Graph A
High School GPA below 2.5
in college preparatory courses



(A)

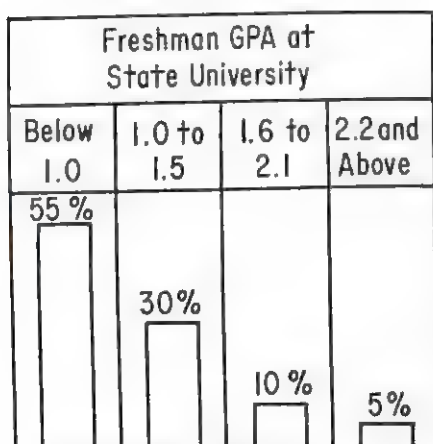
Graph B
High School GPA 2.5 and above
in college preparatory courses



(B)

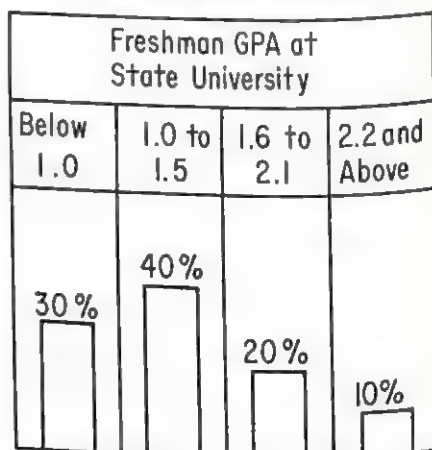
Graphs for Students Scoring in Stanines 6 and 7 (IQ's 104-119)^a

Graph A
High School GPA below 2.5
in college preparatory courses



(A)

Graph B
High School GPA 2.5 and above
in college preparatory courses



(B)

Fig. 17.1 Graphs (based on local expectancy tables) for Use in Interpreting Data on *Both* Scholastic Aptitude and High School Grades, as a Basis for Predicting Achievement at the State University.

NOTE TO STUDENTS: This graph shows how former students of this school, whose average scores on two scholastic aptitude tests were similar to yours, achieved during their fresh-

man year at the state university. Use your own grade-point average in college-preparatory subjects to decide whether you should use Graph A or Graph B. If your grade-point average is approximately 2.5, look up your expectancies in each graph, and average them.

These graphs are based on the same general approach as those in George E. McCabe, "Test Interpretation in the High School Guidance Program," *Test Service Bulletin No. 93* (New York: Harcourt, Brace & World, Inc., n. d.). Single copies are free on request to the Division of Test Research.

- These headings would *not* appear on the student's form; instead a code number should be placed in one corner of the form to facilitate the counselor's using the right chart with each student.
-

Testing the Suitability of a Decision Already Made by an Individual

Some counselors are delegated responsibility for judging the suitability of specific educational and vocational plans; for example, a counselor employed by the Veterans' Administration is required to pass on the appropriateness of vocational plans for trainees. Many other counselors are faced with similar problems, in that a counselee seeks their aid in assessing the feasibility of a specific choice of college or vocation. That is, the counselor is asked to predict what a college admissions officer or an employer might decide. In such situations, the counselor who has at hand expectancy tables based on an adequate number of cases and relevant to the specific situation can interpret the student's test data, in combination with his grades, to help him formulate hypotheses about the probable results of a specific choice. Table 17.1 illustrates a useful expectancy table based on a single predictor variable. Figure 17.1 introduces a second nontest variable and yet keeps the presentation easy to understand. The use of prediction equations or profile similarity studies, discussed in a later section, might prove even more useful in considering such questions.

Choosing the Optimum Level of Work

A decision concerning how heavy a load of college-preparatory subjects should be taken (as in Problem 6 of Chapter 1), or how selective a college to attend, involves estimating the student's *level* of general academic aptitude, and past achievement, as well as his level of motivation and self-discipline as a student. Here again an expectancy table such as Figure 17.1 would be helpful in estimating the student's readiness for a moderate or heavy load of college-preparatory work. Such data, however, must be supplemented by nontest data on the student's study habits, level of aspiration, and the like.

Expectancy tables, which help a student to see that achieving a B grade

in algebra appears possible but not probable for him, might help him to postpone his college preparatory work in foreign language and concentrate his study time on a subject that appears to be fairly difficult for him. Then, when his year's experience in algebra has provided additional data, he would be in a better position to make further plans.

It is often helpful to both students and parents to avoid overemphasis on the ninth grade as a crucial choice point with respect to college-preparatory vs noncollege-preparatory courses. Although present decisions should be made in terms of the *perspective of a long-range plan*, a maximum of flexibility for choosing alternatives at later choice points should be retained whenever possible. Educational and vocational planning involves a series of decisions; the road chosen at a particular choice point may have many branches.

Choosing among Alternatives on the Basis of Intraindividual Differences

When a student is attempting to choose *among* several vocations, and he has an adequate level of general ability for each of them, *differential prediction* is involved. On the basis of his intraindividual differences, we want to predict whether he will be more successful and happy in one vocation than in another. As was emphasized in the discussion of Problem 7 in Chapter 1, tests that predict a student's level of achievement in academic work may be of little value in predicting whether he would do *better* in nursing or teaching, or whether he would find *greater* job satisfaction in one vocation than the other.

In helping students make hypotheses involving differential prediction, the counselor tends to use a clinical approach to the study of test and nontest data. Expectancy charts can help him in checking the suitability of a student's tentative decision or in estimating his probability of success in colleges or curricula demanding different levels of academic aptitude. Published research studies to assist in making *differential predictions* with respect to the student's relative chances of success in various fields are few and inadequate.

The counselor can determine whether the student's *interests* are more nearly similar to persons successfully engaged in one occupation or the other. He may be able to report the probability of his gaining entrance into, and completing, the training programs for two or more different fields. But in so far as predicting the difference in his ultimate "success" in two or more vocations, research data provide no adequate basis for prediction. Fortunately, there is quite a variety of jobs *within* many vocational areas so that persons often find positions that are satisfying to them even though the general vocational field was a far-from-ideal choice.

DIFFERENTIAL PREDICTION BY MEANS OF PREDICTION EQUATIONS Some schools and colleges have obtained local data that help in differential prediction; for example, they may have developed prediction equations by which they can predict a student's grade-point average in each of several curricula on the basis of their entrance-test data. When these equations involve two or more predictor tests, they are called *multiple regression equations*. These are similar in type to the prediction or regression equations discussed in Chapter 3 but involve an optimally weighted combination of student's scores on two, three, or more predictor tests.¹¹

In order to obtain a better understanding of multiple-regression equations, we will use a simplified illustration. Let us assume that we have available students' scores on several tests, which might be useful in predicting whether or not they are likely to succeed in ninth-grade algebra. In making our predictions, we would like to take into account both the most recent intelligence test (variable 1) and one other test that would contribute most to accuracy of prediction.

We will call the predicted variable (grade in algebra) variable y and the predictor tests will be numbered 1, 2, 3, 4, and 5. Let us assume that each of the five predictor tests correlates .50 with grade in algebra.

Variable 1 (8th grade IQ test)	$r_{y1} = .50$
Variable 2 (arithmetic computation test)	$r_{y2} = .50$
Variable 3 (arithmetic reasoning test)	$r_{y3} = .50$
Variable 4 (reading test)	$r_{y4} = .50$
Variable 5 (6th grade IQ test)	$r_{y5} = .50$

Although this example has been artificially simplified, these are all variables that might well show an r of about this size with grade in algebra. These tests differ, however, in the extent to which they provide new information. The correlations of each of the four tests (2 through 5) with variable 1 (the most recent intelligence test) differ as follows:

$r_{12} = .30$ (IQ and arithmetic computation)
$r_{13} = .50$ (IQ and arithmetic reasoning)
$r_{14} = .70$ (IQ and reading)
$r_{15} = .80$ (IQ and previous IQ test)

¹¹ Ordinarily, a test of scholastic aptitude is the first predictor test included. A second, and sometimes a third, predictor test are selected on the basis of relatively high r 's with the criterion (for example, grade-point average in a specific major field) and relatively low r 's with the other predictor tests. In interpreting predicted grade-point averages in different major fields, the counselor must make allowance for departmental differences in grading standards through the use of some type of converted score.

We can now compute R , the multiple correlation¹² between y and each pair of variables. In each pair, we combine the most recent intelligence test with each of the other predictor tests.

$$\begin{aligned} R_{y,12} &= .62 \text{ (algebra grade with IQ and arithmetic computation)} \\ R_{y,13} &= .58 \text{ (algebra grade with IQ and arithmetic reasoning)} \\ R_{y,14} &= .54 \text{ (algebra grade with IQ and reading)} \\ R_{y,15} &= .53 \text{ (algebra grade with both recent and 6th grade IQ tests)} \end{aligned}$$

According to these results, if we can use only two tests in prediction, the best test to select from this group would be test 2 on arithmetic computation. The use of the arithmetic computation test increases the correlation with algebra grade from .50 to .62 because this test has the lowest r with test 1 and the smallest percentage of overlapping variance.

This test provides more new information than any other single test. Whether or not a counselor uses prediction equations, he needs to realize that both he and his counselee can keep only a limited number of factors in mind at a time. Hence, in his selection of tests to be administered or in his interpretation of test data relevant to a decision, he should choose tests that are highly related to success on the criterion and that have a relatively low relationship with each other. If we were going to prepare a two-variable expectancy chart for students, we would choose the arithmetic computation test as a second variable in preference to other tests that overlap to a greater degree with the most recently administered intelligence test.

We discussed in Chapter 3 the simple regression equation in standard-score form

$$\bar{z}_y = r z_x$$

where the coefficient of z_x was r , the slope of the prediction line. When three variables are involved in a correlation, the equation for the prediction line in three-dimensional space is a little more complex.

$$\bar{z}_y = \beta_1 z_1 + \beta_2 z_2$$

where z_1 and z_2 represent the standard score on the first and second predictor tests (in this case intelligence and arithmetic computation) and β_1

¹² The procedures for computing R , the multiple correlation between a predicted (dependent) variable and one or more predictor (independent) variables are given in standard textbooks on statistics, for example, Quinn McNemar, *Psychological Statistics* (New York: John Wiley and Sons, Inc., 1962), pp. 174–187. If a person wants to obtain approximate values of a multiple R involving only three variables, a nomograph can be used. Frederic M. Lord, "Nomograph for Computing Multiple Correlation Coefficients," *Journal of the American Statistical Association*, vol. 50 (December 1955), pp. 1073–1077.

and β_2 (known as *beta*, or multiple-regression, coefficients) represent the weights that must be assigned the first and second predictor tests in order to achieve the most accurate prediction.¹³

In this illustrative problem, in which standard scores are used and both tests had the same correlation with the criterion, the *beta* weights would be equal. However, if variable 1 (8th-grade intelligence test) had correlated .60 with success in algebra and variable 2 (the arithmetic computation test) had correlated only .40, the intelligence-test standard scores would have been given more than twice the weight of the arithmetic computation scores in the prediction formula, which would be as follows:

$$\bar{z}_y = .53 z_1 + .24 z_2$$

For a student with a *z* score of +1 in IQ and -1 in arithmetic computation, the predicted *z*-score for algebra grade would be .29, or a percentile rank of 62. However, for a student with the reverse pattern (a *z*-score of -1 in IQ and +1 in arithmetic computation), the predicted *z*-score would be only -.29, or a percentile rank of only 38. The variable having the higher correlation with the criterion is given the greater weight in the prediction formula.

Now that computers are being used to an increasing degree in providing services to school systems, counselors should have made available to them students' predicted grade-point-averages in different subjects. This type of information would not only be helpful in guidance but would constitute a better basis for identifying "underachieving pupils" than the all-too-common plan of calling in all students with low grades, regardless of their aptitude for the subject field.

The most accurate, objective basis for replying to questions regarding the curriculum or the vocation in which a student would do *best* is provided by substituting this student's scores in a series of multiple regression equations. In this way, we combine relevant data, weighted so as to have optimum predictive validity. The predicted scores we would obtain for each field could be easily compared. The difficulty is that although we have the statistical techniques for such prediction, and fairly adequate tests of many abilities, we do not have adequate data from longitudinal studies on the relationship of predictor test scores and ultimate criteria of success in different educational and vocational fields. Hence, for a long time to come, differential prediction in *vocational* guidance will involve the subjective interpretation of test and nontest data, utilizing such techniques as interpretation of profiles.

¹³ The formulas for *beta* coefficients in three-variable problems, and problems involving more than three variables, are given in standard textbooks on statistics.

STUDYING PROFILE SIMILARITY The publishers of aptitude test batteries (such as the *Differential Aptitude Tests*) and of interest inventories (such as the *Kuder Preference Record*) have published profiles of mean scores on different subtests for a number of occupational groups. The counselor using such profiles should ascertain the number of cases on which the occupational profile is based and whether the sample appears to be a representative one.

The publishers of a few achievement tests have also prepared profiles of means, to be used in student counseling. For example, the publishers of the *Iowa Tests of Educational Development* have prepared profiles of mean subtest scores for students, tested in high school and later graduating with college majors in 11 different fields. A student's profile can be compared with profiles for each of the 11 curricula to see which one it most resembles. Or, if the student has two or three tentative choices, his profile can be checked against profiles of mean subtest scores for each curriculum.

Profiles cannot be as easily or objectively compared as the scores obtained from regression equations. Visual inspection, that is, comparing the student's profile with each of several occupational profiles to get a general impression of similarity, is not very satisfactory.

One can compute the *D*-statistic¹⁴ for each of the fields being considered, that is, the square root of the sum of all squared differences between student A's interest or ability scores and the mean scores for each of the relevant occupational profiles.

In Table 17.2, we have compared Donna's scores on the DAT with the occupational profiles for nurses and teachers. Donna's standard scores are very similar to the mean scores for nurses in several tests: NA (numerical ability), MR (mechanical reasoning), CSA (clerical speed and accuracy), and both the language usage tests. However, we need data from other sources before we can conclude whether these similarities will contribute to her success in nursing. The only sizable differences are Donna's moderate superiority to the nurses' average in both verbal reasoning and abstract reasoning. Certainly, there is nothing in this comparison of profiles that would discourage Donna from choosing nursing if the interest

¹⁴ The *D*-statistic represents the distance between two points: (1) the point represented by graphing the mean scores for an occupational group and (2) the point represented by graphing the corresponding scores for an individual. If two predictor tests are used, these are points in 2-dimensional space; for three variables they are points in 3-dimensional space, and the like. For an illustration of the use of the *D*-statistic, an explanation of its meaning and a comparison with the linear discriminant function, which should be used when tests show substantial intercorrelation, the reader is referred to Jum C. Nunnally, Jr., *Tests and Measurements: Assessment and Prediction* (New York: McGraw-Hill Book Company, Inc., 1959), pp. 129-134.

Table 17.2

An Illustration of Computation of the *D*-Statistic for Subtests of the Differential Aptitude Test^a

SUBTEST OF DAT	Mean standard scores for		Standard scores for Donna	Comparison with nurses		Comparison with teachers	
	NURSES	WOMEN TEACHERS		DIFF.	DIFF. ²	DIFF.	DIFF. ²
VR—verbal reasoning	.77	.87	1.00	+.23	.0529	+.13	.0169
NA—numerical ability	.67	1.00	.75	+.08	.0064	— .25	.0625
AR—abstract reasoning	.60	.87	.92	+.32	.1024	+.05	.0025
SR—space relations	.73	.63	.90	+.17	.0289	+.27	.0729
MR—mechanical reasoning	.35	.55	.20	— .15	.0225	— .35	.1225
CSA—clerical speed and accuracy	.20	.60	.30	+.10	.0100	— .30	.0900
LU-I—language usage; spelling	.53	.58	.65	+.12	.0144	+.07	.0049
LU-II—language usage; sentences	.40	.90	.45	+.05	.0025	— .45	.2025
Average	.53	.75	.65		.2400		.5747
<i>D</i> —statistic ^b					.49		.76

^a The standard scores for the samples of nurses and women teachers were obtained by translating the mean percentile ranks provided in the following bulletin into z-score equivalents on the normal curve. Data from "The D.A.T.—A Seven-Year Follow-up," *Test Service Bulletin* No. 49 (New York: The Psychological Corporation, 1955), p. 13.

^b The *D*-statistic is the square root of the sum of the squared differences between the individual's subtest scores and the corresponding mean scores of the group with which he is being compared.

inventory results and relevant nontest data seem to support this vocational choice.

When Donna's data are compared with the profile of mean scores for women teachers, we note similarity with respect to VR, AR, and LU-I. As compared with teachers, Donna has relatively low scores in NA, MR, CSA, and LU-II (language usage, sentences). Donna's score in NA is well above the average for students in general, but it would be helpful to know her percentile rank *within* the teacher group. Because there are a wide variety

of opportunities within teaching, however, and some require less NA ability than others, this subtest score would not, in itself, discourage consideration of teaching. Donna's lower score on MR would seem to offer no cause for concern. Her low score on CSA *might* be attributable to an emphasis on working accurately, which lowers the scores of some conscientious students on this highly speeded test. Donna's most disconcerting score is that in Language Usage II. Fortunately, this is an area in which she could markedly improve her ability if she were highly motivated.

Computation of the *D*-statistics indicates that Donna's profile resembles that for nurses more than that for teachers. One realizes from this example, however, some of the disadvantages of routine use of the *D*-statistic for comparisons of profiles. *Equal weight* is given to differences on all tests, although some differences are surely more significant than others. Professional judgment needs be introduced into the interpretation process as we have done in the preceding paragraphs.

If the *SD*'s of scores for each occupational group are given, one could apply the *D*-statistic only to those subtests with (1) relatively high mean scores and (2) relatively small *SD*'s. In this way, we would save work by eliminating from consideration tests in which most students could meet occupational requirements and also tests in which members of the occupation show such a wide spread of scores that differences in this variable may be assumed not to be highly relevant to success.

Since the *D*-statistic fails to take into account the relative weight that should be assigned different factors, it is inferior to the multiple-regression equations in predicting success from ability test scores, but may prove superior for factors that fail to show a linear relationship to success, such as interests or personality characteristics. That is, profile comparison is especially useful in judging the appropriateness of interests or personality characteristics for different vocations. Scores on these inventories often show a curvilinear relationship to success (for example, there may be an optimum range of score values in certain personality traits, with both higher and lower scores being less predictive of success). Multiple regression equations, which assume a linear relationship¹⁵ cannot be used in such cases.

In all interpretations of profiles for individual students, we must be careful not to read too much significance into differences that do not exceed the standard error. The student should review the section on the standard errors of differences in Chapter 3. One of the best approaches is to incorporate the concept of measurement error into the profile itself, as in the profile for the *Differential Aptitude Tests* shown in Chapter 6.

In interpreting profiles for the *Kuder Preference Record* and other tests

¹⁵ In a linear relationship, the higher the predictor scores, the higher the success on the criterion.

of the forced-choice type, the counselor needs to keep in mind that the scores are based on the individual's expression of *preferences*, not on any absolute measure of his *degree of interest in a field*. Hence, everyone's profile is centered around the median. There are no profiles that *average high* or *average low*; it is impossible for the person with high enthusiasm for many areas to show an elevated score in all of them.

When we use aptitude-test profiles to help students in vocational planning, we are not only concerned about the reliability of differences between subtests in the student's profile. We are also concerned with whether these differences represent a fairly stable pattern of intraindividual differences. Unless the pattern has stability over time, a counselor of ninth graders has little justification for using it in a discussion of post-high-school plans. In a research study, in which ninth-grade *differences* between DAT subtests were correlated with twelfth-grade differences, the *r*'s ranged from a low of .20 (numerical minus abstract) to .74 (mechanical minus spelling). The reader will recognize that the second pair of tests mentioned would have a much lower intercorrelation than the first pair. According to the findings of this study, differences among mechanical, clerical, and the over-all level of the verbal-language-numerical tests showed sufficient stability over time to justify interpretation. It appeared doubtful that predictions based on differences *between* verbal, numerical, and abstract scores were predictive of later differences.¹⁶ Although this study was made of the DAT, it raises questions about the justifiability of assuming the stability of profile differences on any test batteries that show fairly high intercorrelations among the tests.

USE OF CRITICAL OR CUT-OFF SCORES Research studies that would provide counselors with data regarding the critical minimum scores on various aptitudes for different vocations are urgently needed. This approach has been used on the GATB (*General Aptitude Test Battery*); that is, critical scores are provided for occupational families (groups of related occupations) for each of three aptitudes that seem to be most important for that occupational group. The critical scores, on the average, correspond to the 33d percentile of workers in the occupations involved. Perhaps the most efficient approach to differential prediction from ability scores is to use cut-off scores to reject occupations or curricula for which the student or employee does not reach a minimum score in critical abilities. Then, for curricula or vocations that still seem feasible, one would use multiple-regression equations for predicting success from ability test scores.

¹⁶ Jerome E. Doppelt and George K. Bennett, "A Longitudinal Study of the Differential Aptitude Tests," *Educational and Psychological Measurement*, vol. 11 (Summer 1951), pp. 228-237.

COMBINING GROUP AND INDIVIDUAL APPROACHES IN HELPING HIGH SCHOOL STUDENTS IN SELF-APPRAISAL AND LIFE PLANNING

Since the interpretation of test data to students affects their self-concepts and may produce conflicts, every effort should be made to individualize the program of test interpretation. However, it is only realistic to realize that extensive programs of aiding students in self-appraisal can usually be undertaken only if group approaches are used in at least certain aspects of the work. Hence, we will attempt to illustrate how a combination of individual and group approaches might be used in a large-scale program of interpreting test data to students who are studying a unit in life planning. The remainder of this chapter section is a report of an actual experience in the interpretation of test data to an 11th-grade class completing a vocational-guidance unit under a core-class teacher. The various student experiences that are described took place at intervals over a period of eight weeks.

Developing General Concepts Basic to Student Self-Appraisal

Group approaches can be used in preparing students for individual conferences regarding their own test data. In this 11th-grade class, a short film ("Pups and Puzzles") was shown, emphasizing how tests of abilities, personality traits, and the like are used by personnel directors in modern industrial plants—as a basis for placing applicants in jobs in which they can work most effectively, rather than as hurdles to eliminate applicants. On the basis of class discussion of the film and home study on several pamphlets on self-appraisal,¹⁷ basic principles regarding the values and limitations of test data were formulated and listed on the board.

All types of test and nontest data that would help in self-appraisal were suggested and listed on the chalkboard. In turn, the values and limitations of achievement, aptitude, and interest tests were considered as well as sources of evidence that might help the student to corroborate or question his own test data. Significant unmeasured aptitudes and other unmeasured factors affecting vocational planning were listed and their relative importance considered.

¹⁷ For example, such pamphlets as *Discovering Your Real Interests* (Chicago: Science Research Associates, 1961); *You and Your Mental Abilities* (Chicago: Science Research Associates, 1959).

Studying the Significance of Test Data for Problems of Vocational Choice

Copies of the tests students had taken were distributed to remind them of their content and organization. Illustrative profiles for hypothetical students (using the mimeographed form reproduced in Figure 17.2) were then distributed and described. Questions were encouraged and fully discussed. Typical problems (such as the flat or poorly differentiated interest profile, discrepancy between interest and aptitude patterns, and the like) were illustrated by these profiles. In the discussion, emphasis was placed on the results of achievement and interest tests, as well as tests of such special aptitudes as clerical, spatial, and mechanical, in which a low score was not threatening to students. There was a brief discussion of the significance of the VR and NA tests of the DAT as being indicative of readiness for college-preparatory courses. It was emphasized that some colleges admit only students with high scores on these or similar tests while other colleges provide a wide variety of courses for students with different ability patterns.

The class then discussed the ability and personality requirements of various occupations (or occupational families). Sources of occupational information and means of evaluating them;¹⁸ were discussed prior to a trip to the school library and an explanation about their file of pamphlets on occupations. Students discussed how to organize their self-appraisal and occupational-information data for comparison in a paper on vocations.¹⁹

Preparing Profiles of Test Data for Individual Students

One of the most time-consuming tasks involved in a self-appraisal program is the preparation of individual profiles for each student. In the group-guidance class described above, each student, under teacher supervision, (1) wrote in his name and other personal data on three copies of his test profiles; (2) recorded raw scores on *all* tests; (3) recorded percentile ranks on tests for which they were available, that is, achievement tests and interest inventories but *not* aptitude tests; and (4) graphed percentile ranks for achievement tests and interest inventories. The profiles were then returned to the teacher.

¹⁸ Max F. Baer and Edward C. Roeber, *Occupational Information: Its Nature and Use* (Chicago: Science Research Associates, 1951).

¹⁹ As an aid to students in interpreting test and nontest data about their abilities, interests, and values and using such data in educational and vocational planning, Katz has developed a work text for eight- or ninth-graders, which contains self-appraisal charts and questionnaires, plus explanations in the teen-ager's own language. Martin R. Katz, *You, Today and Tomorrow* (Princeton, N. J.: Educational Testing Service, 1955). An accompanying *Teacher's Guide* is also available.

Preparing for Individual Interviews Concerning Self-Appraisal and Vocational Choice

In preparation for each individual conference, the guidance teacher studied both the student's cumulative record and the profile of recent test data. He was alert to discrepancies between recent and earlier sets of data and other questions raised by the cumulated test findings. Converted scores for the aptitude tests were added to the profiles for most of the students. However, with students for whom such a presentation might be threatening, aptitude-test data were discussed on the basis of the guidance teacher's interpretation of raw scores whenever he felt it advisable to introduce such data into the conference.

During the individual conferences, the teacher encouraged the student to react to his scores and interpret their meaning for him, urged him to bring in and relate to the test data other relevant sources of information, and tried to correct any misconceptions evidenced in the student's written interpretation of the data.

Analysis of an Illustrative Student Profile of Test Data

Figure 17.2 presents the profile of test data for Dick Jones, an 11th-grade student. At the time that deficiency notices were sent out, Dick had received a failure notice in geometry and an unsatisfactory notice (D grade) in Spanish. His work in the English-social studies class, which he took with his guidance teacher, had been consistently below average; he frequently failed to hand in assignments and seemed to daydream and dawdle during supervised study periods rather than attacking his work systematically.

Dick was such a fine-looking young man and came from such a good family that Mr. Peters had attributed his low achievement merely to the indifference or actual dislike that many boys of his age show toward English and social studies. A number of times he had seen the boy working surreptitiously on geometry when he should have been reading a literature assignment; hence, he had tended to group Dick in his own mind with the other boys in the class who were planning to be engineers and who showed a marked preference for mathematics and science as compared with English and social studies.

Dick's two deficiency notices, however, helped Mr. Peters to realize that he had no objective basis for his generalizations. Observation of Dick's behavior in class revealed poor work habits and some resistance to adult-imposed assignments. Study of his cumulative record revealed that his family lived in a very high socioeconomic neighborhood and that Dick was the younger of two children, the other child being a girl five years older who was totally deaf.

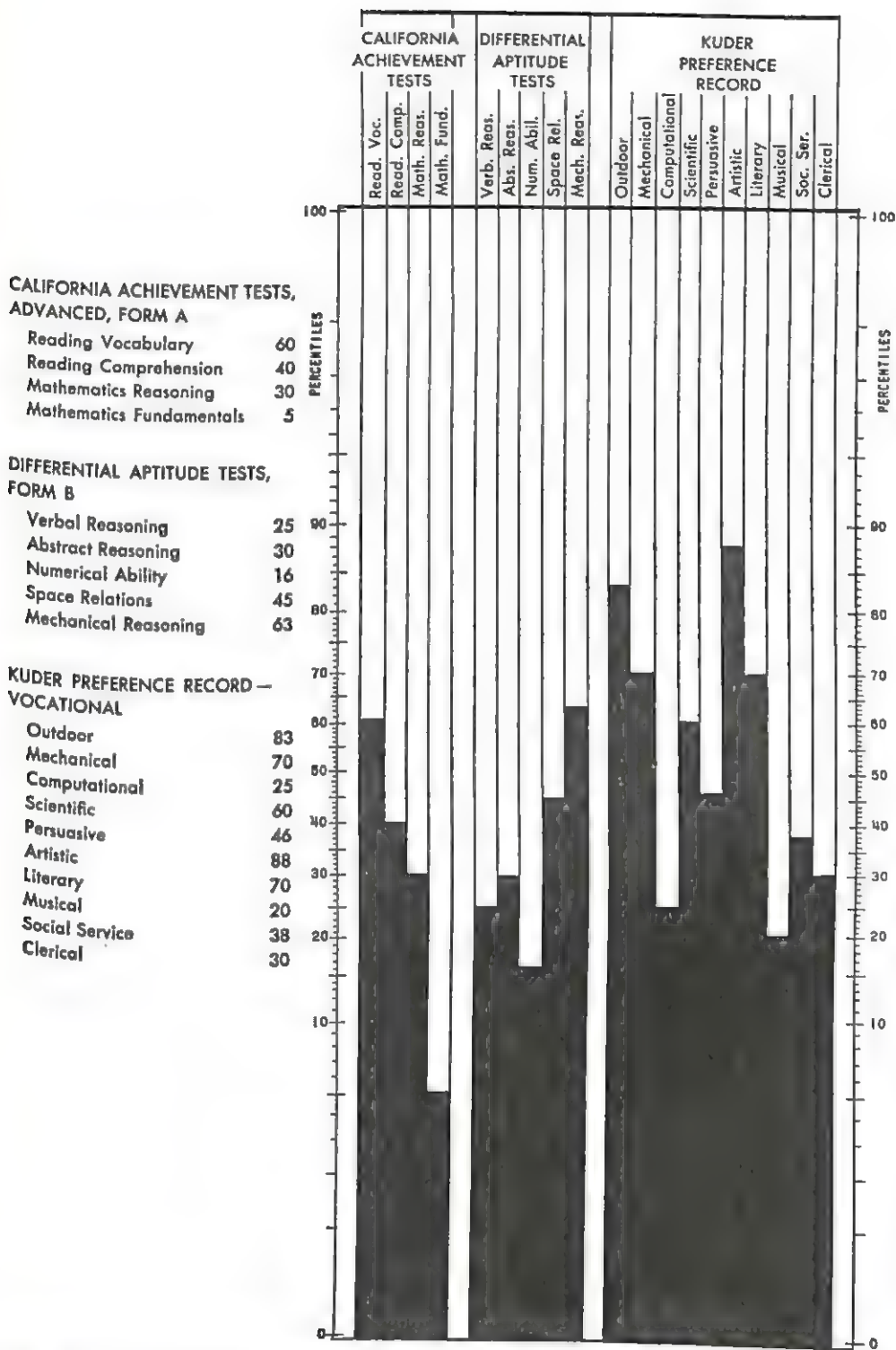


Fig. 17.2 Percentile Ranks and Profile of "Dick Jones" on Three Test Batteries.

An interview with the mother confirmed the teacher's hypothesis that the parents' frustration concerning the handicap of the older child plus their desire to maintain the socioeconomic status of the family combined to make it seem imperative that Dick enter college and prepare for a profession. Dick did not question the desirability of this goal, but was having a very disheartening experience in trying to realize his parents' ambitions for him. Like many students, Dick felt that he was being exploited by his parents because of their almost exclusive emphasis on grades; he felt insecure because he was disappointing them in a crucially important way; he was building up a self-concept of himself as either "dumb" or somehow unable to study and concentrate.

Examination of his achievement-test data indicates a reading-vocabulary score somewhat above average (as would be expected from his superior home environment) but reading-comprehension and mathematics-reasoning scores that represent below-average achievement for his grade level. Especially low was his percentile rank of only 5 in mathematics fundamentals—a score that seems consistent with his low computational interest on the Kuder. Dick had already graphed the achievement- and interest-test data; he had noticed that they were consistent with his failing grade in mathematics.

In the conference, the guidance teacher revealed that Dick's percentile rank in numerical ability (DAT) was low and commended him, in view of this rank, for having earned a C in algebra in the tenth grade and a B in general mathematics in the ninth grade. These marks showed good effort in fields that were inherently difficult and uninteresting for him.

Dick soon recognized that his tentative vocational goal of engineering was highly inadvisable, especially when the teacher explained that his numerical ability score would be in the lowest 1 percent as compared with engineering-school freshmen.

Dick's attention was called to the fact that his score in space relations was average, and that his score in mechanical reasoning exceeded approximately two thirds of 11th-grade boys. In two subtests of the DAT most closely related to success in college (verbal reasoning and abstract reasoning), Dick's scores placed him in the lowest quarter and third, respectively, of his class.

The need for considering other vocational possibilities was apparent to Dick. It seemed evident from his family background and the aspirations of his parents that certain seemingly logical possibilities (such as auto mechanic) were almost out of the question at the present time. Examination of the interest-inventory profile revealed high interests in the mechanical and artistic areas, which, combined with his high aptitude in mechanical reasoning and high achievement in art courses, suggested industrial design as a possible vocational choice. The combination of high outdoor and

artistic interests, together with Dick's fine appearance and social acceptability, suggested the possibilities of scout leadership, provided that he could obtain the necessary preparation at a college with relatively low entrance requirements. However, Dick's relatively low interest in social service, confirmed by his passivity in social relationships, were contraindications to this vocational choice.

Another possible vocational choice in a field taught at the local junior college was merchandising. This choice seemed feasible in view of Dick's high artistic interest and achievement, average persuasive interest, such important unmeasured factors as fine personal appearance, social acceptability, home background, and availability of training, and the fact that the local demand for young men far exceeded the supply. His artistic interest and ability (as evidenced by his achievement in design courses) suggested that he specialize in window display. Possibilities for rapid promotion of men to administrative positions and the presence in the community of several large department stores seemed to support the advisability of this field as a tentative vocational choice.

Dick seemed greatly relieved to know that he should probably not attempt a four-year program for a college degree and to contemplate his above-average abilities in art and mechanical reasoning. He was delighted to learn of vocational fields in which he might succeed and that would meet his status needs and those of his family. Several next steps were indicated—(1) active exploration (through reading, interviews, and the like) of the vocations suggested above and other new possibilities, (2) enrollment in such exploratory courses as salesmanship or machine shop, and (3) a conference with the parents concerning the interpretation of the test data and such revised plans as Dick would formulate.

SUMMARY STATEMENT

The guidance functions usually performed by specially trained personnel and those usually performed by teachers were reviewed. The best division of responsibilities for guidance varies with the size of the school, the specialized staff and facilities available, the attitudes and training of its administrative staff members, and many other factors.

Four contentions of guidance leaders who minimize the importance of test data in counseling were analyzed: (1) that the influence of unmeasured or qualitative variables is so great that the measurable characteristics, in comparison, are relatively insignificant; (2) that existing tests have serious limitations; (3) that the student is likely to take a passive, dependent role in the conference regarding test data; and (4) that interpretation of test results to a student may threaten his concept of self and adversely affect the counseling relationship.

The following principles should be observed in the use and interpretation of test data in guidance: (1) the best available tests for the purpose should be

used; (2) responsibility for the interpretation of test data should be placed with persons who are competent to handle those responsibilities; (3) test data should be considered in the context of all other available information; (4) test results and other evaluation data should be interpreted in terms of probabilities, rather than certainties; (5) evaluation data should be brought into the counseling interview as they help in meeting a need and should be presented objectively and impersonally, with the student being encouraged to make his own interpretations of their personal meaning for him and to express his reactions; (6) guidance workers must interpret with special caution data from all tests on which the examinee can vary his response at will; (7) the counselor's approach should be conditioned by his realization that the student's interpretation of test data is sometimes more of an emotional than a rational process.

As an illustration of the use of test data in counseling, two of the hypothetical problems presented in Chapter 1 were next considered. The value of expectancy tables in judging the suitability of choices already made or in choosing the optimum level of work is emphasized. Three different approaches which are helpful in differential prediction (choosing among alternatives on the basis of intraindividual differences) were studied: (1) using prediction or regression equations, (2) studying profile similarity and (3) using critical or cut-off scores.

Group and individual approaches can be effectively combined in helping high school students to use test data in self-appraisal and life planning. An illustrative application was described of the combined use of these approaches in interpreting data to an eleventh-grade class.

SELECTED REFERENCES

- BIXLER, RAY H., AND VIRGINIA H. BIXLER, "Test Interpretation in Vocational Counseling," *Educational and Psychological Measurement*, vol. 6 (Spring 1946), pp. 145-155.
- BORDIN, EDWARD S., "The Implications of Client Expectations for the Counseling Process," *Journal of Counseling Psychology*, vol. 2 (1955), pp. 17-21.
- , "Four Uses for Psychological Tests in Counseling," *Educational and Psychological Measurement*, vol. 11 (Winter 1951), pp. 779-781.
- COLLEGE ENTRANCE EXAMINATION BOARD, *Manual of Freshman Class Profiles*. New York: The Board, 1963.
- CRAVEN, E. C., *The Use of Interest Inventories in Counseling*. Professional Guidance Series. Chicago: Science Research Associates, 1961.
- CRONBACH, L. J., AND G. C. GLESER, "Assessing Similarity between Profiles," *Psychological Bulletin*, vol. 50 (November 1953), pp. 456-473.
- DOBBIN, JOHN E., *Bridging the Gap in Guidance*. Princeton, N.J.: Educational Testing Service, 1962.
- FROELICH, CLIFFORD P., AND K. B. HOYT, *Guidance Testing and Other Student Appraisal Procedures for Students and Counselors*. Chicago: Science Research Associates, 1959.
- GOLDMAN, LEO, *Using Tests in Counseling*. New York: Appleton-Century-Crofts, 1961, 434 pp.
- HILLS, JOHN R., AND OTHERS, "Admissions and Guidance Research in the University System of Georgia," *Personnel and Guidance Journal*, vol. 39 (February 1961), pp. 452-457.

- HOPKE, WILLIAM, "Getting Guidance Information into the Hands of Teachers," *School Counselor*, vol. 9 (December 1961), pp. 62-65.
- MOSIER, CHARLES L., "Batteries and Profiles," in E. F. Lindquist, ed., *Educational Measurement*. Washington, D.C.: American Council on Education, 1951.
- ROTHNEY, JOHN W. M., AND BERT A. ROENS, *Counseling the Individual Student*. New York: Holt, Rinehart and Winston, Inc., 1949.
- SEGEL, DAVID, AND OTHERS, *An Approach to Individual Analysis in Educational and Vocational Guidance*. Bulletin, 1959, No. 1. United States Office of Education. Washington, D.C.: Government Printing Office, 1959.
- SUPER, DONALD E., "The Critical Ninth Grade: Vocational Choice or Vocational Exploration," *Personnel and Guidance Journal*, vol. 39 (October 1960), pp. 107-109.
- , "Vocational Adjustment: Implementing a Self-Concept," *Occupations*, vol. 30 (November 1951), pp. 88-92.
- TYLER, LEONA E., *The Work of the Counselor*, 2d ed. New York: Appleton-Century-Crofts, 1961.
- Using the Iowa Tests of Educational Development for College Planning*. Chicago: Science Research Associates, 1957.

DISCUSSION QUESTIONS AND SUGGESTED ACTIVITIES

1. Give two or more illustrations of counseling situations in which expectancy tables could be used to advantage.
2. What additional data would be desirable in counseling Mary (the girl whose DAT aptitude test profile is shown in Chapter 6)? In counseling Dick (the boy whose test data were discussed in this chapter)?
3. What procedures might a counselor use to minimize (a) the risk that the counselee would passively accept test data as authoritative and not to be questioned, and (b) the risk that the student's self-concept might be threatened by the interpretation of test results.
4. In what types of counseling situations is there need to make differential predictions?
5. Comment on the advantages and disadvantages of using regression equations and profile analysis in problems requiring differential prediction.
6. What data would be useful to the counselor in counseling students on the advisability of electing a shorthand course?

Appendixes

APPENDIXES

Appendix A

A Selected List of Standardized Tests for the Elementary and Secondary Schools

TABLE I	Achievement Test Batteries
TABLE II	Achievement: Business Education
TABLE III	Achievement: Foreign Language
TABLE IV	Achievement: Health Education
TABLE V	Achievement: Language and Literature
TABLE VI	Achievement: Mathematics
TABLE VII	Achievement: Music
TABLE VIII	Achievement: Reading and Vocabulary
TABLE IX	Achievement: Science
TABLE X	Achievement: Social Studies
TABLE XI	Aptitude: Group Tests of General Mental Ability
TABLE XII	Aptitude: Individual Tests of General Mental Ability
TABLE XIII	Aptitude Test Batteries
TABLE XIV	Aptitude: Art
TABLE XV	Aptitude: Business-Clerical
TABLE XVI	Aptitude: Foreign Language
TABLE XVII	Aptitude: Manual Dexterity and Mechanical Aptitude
TABLE XVIII	Aptitude: Mathematics
TABLE XIX	Aptitude: Music
TABLE XX	Aptitude: Reading Readiness
TABLE XXI	Aptitude: Science
TABLE XXII	Interest Inventories
TABLE XXIII	Personal-Social Adjustment
TABLE XXIV	Work-Study Habits and Skills
TABLE XXV	Miscellaneous

Appendix B

Publishers of Standardized Tests

Appendix C

Methods of Expressing Test Scores (Based on the Normal Curve)

Appendix D

Selected Tables

Appendix E

Glossary of Symbols and Terms Used in Measurement
and Evaluation

APPENDIX A

A Selected List of Standardized Tests for Elementary and Secondary Schools

Table 1
Achievement Test Batteries

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN. ^a	NO. OF FORMS	SCORING ^b	PUBLICATION DATES ^c	PUBLISHER	REVIEWS ^d
<i>California Achievement Tests 1957 edition</i>							
(CAT)—E. W. Tiegs and W. W. Clark ^e							
Lower Primary	1-2	90-110	2	H	1934-59	CTB	5-2
Upper Primary	3-4.5	125-145	2	H			
Elementary	4-6	145-165	4	M, Q			
Junior High	7-9	170-190	4	M, Q			
Advanced	9-14	160-180	3	M, Q			
<i>California Tests in Social and Related Sciences (CTSRS)—G. S. Adams and others</i>							
Elementary ^f	4-8	170	2	M, Q	1946-55	CTB	5-4 4-23
Advanced ^g	9-12	170	2	M, Q	1954-55		5-4
<i>Cooperative General Achievement Tests Revised Series (GAT)</i>							
Test 1—Social Studies	12-13	40	2	M	1937-56	ETS	5-787 4-668
Test 2—Natural Sciences	12-13	40	2	M			5-703 4-595

Table 1 (Continued)
Achievement Test Batteries

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN. ^a	NO. OF FORMS	SCORING ^b	PUBLICATION DATES ^c	PUBLISHER	REVIEWS ^d
Test 3—Mathematics	12-13	40	2	M			5-420 4-379
<i>Essential High School Content Battery—</i> D. P. Harry and W. N. Durost ^h	9-13	200-225	2	M	1950-51	HBW	4-9
<i>Iowa Tests of Basic Skills (ITBS)—</i> E. F. Lindquist and A. N. Hieronymous							
Test A—Arithmetic	3-9	60	2	M, Q	1955-56	HM	5-16
Test L—Language Skills	3-9	67	2	M, Q			
Test V & R—Vocabulary and Reading Comprehension	3-9	72	2	M, Q			
Test W—Work-Study Skills	3-9	80	2	M, Q			
<i>Iowa Tests of Educational Development</i> (ITED)	9-13	459	1	M	1942-59 (Norms— 1962)	SRA	5-17 4-17 3-12
Test 1—Understanding of Basic Social Concepts		55	1	M			5-791
Test 2—General Background in the Natural Sciences		60	1	M			5-713
Test 3—Correctness and Appropriateness of Expression		60	1	M			5-197
Test 4—Ability To Do Quantitative Thinking		65	1	M			5-428
Test 5—Ability To Interpret Reading Materials in Social Studies		60	1	M			5-685

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN. ^a	NO. OF FORMS	SCORING ^b	PUBLICATION DATES ^c	PUBLISHER	REVIEWS ^d
Test 6—Ability To Interpret Reading Materials in Natural Sciences		60	1	M			5-686
Test 7—Ability To Interpret Literary Materials		50	1	M			5-217
Test 8—General Vocabulary		22	1	M			5-235
Test 9—Use of Sources of Information		27	1	M			6-692
<i>Metropolitan Achievement Test (MAT)—</i>							
R. D. Allen and others ¹							
Primary 1	1	95-100	3	H	1958-61	HBW	4-18 3-13
Primary 2	2	105-115	3	H			
Elementary	3-4	160-175	3	H			
Intermediate	5-6	250-280	3	M			
Advanced	7-9	260-290	3	M			
<i>Sequential Tests of Educational Progress</i>							
(STEP) ¹							
Level 4	4-6	455	2	M	1957	ETS	5-24
Level 3	7-9	455	2	M			
Level 2	10-12	455	2	M			
Level 1	13-14	455	2	M			
<i>SRA Achievement Series—</i>							
L. Thorpe, and others ^k							
Grades 1-2	1-2	95-125	2	H	1954-59	SRA	5-21
Grades 2-4	2-4	95-125	2	H			
Grades 4-6	4-6	355-445	2	M			
Grades 6-9	6-9	300-375	2	M			
<i>SRA High School Placement Test</i> , ¹ Series AP	8.5-9.5	195	1	Central Scoring Only	1957-61	SRA	5-22

Table I (Continued)
Achievement Test Batteries

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN. ^a	NO. OF FORMS	SCORING ^b	PUBLICATION DATES ^c	PUBLISHER	REVIEWS ^d
<i>Stanford Achievement Test—</i>							
T. L. Kelley and others ^m							
Primary	1.9–3.5	95	5	Q	1923–64	HBW	5-25
Elementary	3.0–4.9	155	5	Q			
Intermediate	5–6	140	5	M, Q			
Advanced	7–9	215	5	M, Q			

^a Where no time limit is specified, approximate working time is given when this information is available.

^b H—hand-scored; M—machine-scored; Q—specially-devised, quick-scoring devices provided to facilitate hand-scoring.

^c The publication dates represent the range of dates for the various editions, forms, and accessories making up the test, as listed in Oscar K. Buros, *Tests in Print* (Highland Park, N. J.: The Gryphon Press, 1961).

^d The number preceding the dash indicates the number of the Buros *Mental Measurements Yearbook*; the number following the dash indicates the test entry. The code "40" is used for the 1940 Yearbook and "38" for the 1938 Yearbook. Reviews of previous editions of a test are included only if the most recent edition has not yet been reviewed in a yearbook. Use of parentheses around an entry number indicates that no evaluative reviews are included; however, information concerning number of forms, testing time, and the like are given, and, in many cases, references to reviews and articles in professional journals.

^e Subtests in reading, language, and arithmetic available as separates.

^f Eight scores in Social Studies I, eight scores in Social Studies II, and seven scores in Related Sciences.

^g Eight scores in American History through the War between the States, eight scores in American History since the War between the States, twelve scores in Related Sciences.

^h Five scores in mathematics, science, social studies, English, total.

ⁱ Subtests in arithmetic, reading, science, and social studies available as separates.

^j Seven subtests at each level: essay test, writing, listening, reading, mathematics, science, social studies.

^k Subtests in language arts, arithmetic, reading, and work-study skills available as separates.

^l Five scores in reasoning, reading, arithmetic, language arts, total, and educational ability.

^m New forms issued annually. Series K prepared for use in Catholic schools also issued annually.

Table II
Achievement: Business Education

NAME OF TEST	GRADE LEVEL	WORKING		NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
		TIME IN MIN.						
<i>Hiatt Simplified Shorthand Test</i> (Gregg)	9-12	50	2	M	1951	KSTC	5-512	
<i>National Business Entrance Testing Program</i>	12-16 Adults	60, 120	3	a	1938-60	NBEA	5-515 (5-506) 3-368	
Bookkeeping Test							(5-508) 3-369	
Business Fundamentals and General Information							(5-511) 3-379	
General Office Clerical Test (including filing)							5-514	
Machine Calculation Test							5-522	
Stenographic Test							5-526	
Typewriting Test								
<i>SRA Typing Adaptability Test</i> — M. Tydlaska and C. White	10-12 Adults	45	1	H	1954-56	SRA	5-518	
<i>SRA Typing Skills</i> — M. W. Richardson and R. A. Pedersen	9-12 Adults	15	2	H	1947	SRA	(3-388d)	
<i>Turse-Durost Shorthand Achievement Test</i> (Gregg)	10-12	50	1	H	1941-42	HBW	(3-392)	

" Two forms are administered only in certified testing centers on specified dates. The completed tests are scored centrally and the results reported later to the schools or employers. Certificates of proficiency are issued to those passing the tests. One form requires 60 minutes

and the second form requires 120 minutes to complete the skills tests. One additional form of the test requiring 120 minutes is available to schools for general testing; no scoring service or proficiency certificates are available for this form.

Table III
Achievement: Foreign Language^a

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
FRENCH							
<i>Cooperative French Test,</i> J. Greenberg, and others	9-16	40	2	M	1932-41	ETS	3-181
<i>Elementary</i> 1-4 semesters (h.s.)							
1-2 semesters (college)	9-16	40	2	M			
<i>Advanced</i>							
<i>Cooperative French Listening</i> <i>Comprehension Test</i> —N. Brooks	9-16	30	2	M	1955	ETS	5-265
LATIN							
<i>Cooperative Latin Test</i> —G. Land							
<i>Elementary</i> 1-4 semesters (h.s.)	9-16	40	2	M	1932-41	ETS	(3-204) 40-1365 38-1065
1-2 semesters (college)							
<i>Advanced</i>	9-16	40	2	M			3-204 38-1064
SPANISH							
<i>Cooperative Inter-American Tests of</i> <i>Language Usage, Spanish Edition</i>	8-13	35	2	M	1963	GTA	

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Cooperative Spanish Test—</i> J. Greenberg, and others							
<i>Elementary</i> 1–4 semesters (h.s.)	9–16	40	2	M	1932–40	ETS	40-1374
1–2 semesters (college)							
<i>Advanced</i>	9–16	40	2	M	1939		40-1373
<i>Spanish-French-German-English Common</i> <i>Concepts Foreign Language Test—</i> B. H. Banathy, and others		40	2	M	1964	CTB	

^a See Chapter 13 for information concerning a series of tests in French, Spanish, German, Italian, and Russian, developed as a cooperative

project of the Modern Language Association of America, the Educational Testing Service, and the United States Office of Education.

Table IV
Achievement: Health Education

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>College Health Knowledge Test, Personal Health</i> —T. H. Dearborn	13-16		1	H	1950-59	STANF	4-478
<i>Health Inventory for High School Students</i> —G. Neher	9-12	40-50	1	M	1942	CTB	3-422
<i>Health Practice Inventory, Revised</i> —E. B. Johns and W. L. Juhnke	10-16	20-30	1	M	1943-52	STANF	5-559
<i>Kilander Health Knowledge Test</i>	9-13	40	2	M	1936-51	HBW	(5-562) (40-1503)

Table V
Achievement: Language and Literature

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
LANGUAGE USAGE AND LANGUAGE SKILLS							
<i>Barrett-Ryan-Schrammel English Test—</i> New Edition	9-13	60	2	M, Q	1938-54	HBW	5-176 40-1267
<i>California Achievement Test: Language—</i> 1957 Edition	(See Achievement Test Batteries)						5-177
<i>Clapp-Young English Test</i>	5-12	25	2	Q	1929	HM	3-117
<i>Cooperative English Test, 1960</i> Revision, Lower Level							
English Expression, Part I Effectiveness of Expression	9-12	40	1	M	1940-60	ETS	5-179
English Expression, Part II Mechanics of Expression	9-12	40	1	M			(5-179) 4-155
Reading Comprehension	9-12	40	1	M			5-645
<i>Cooperative English Test, 1960</i> Revision, Higher Level							
English Expression, Part I Effectiveness of Expression	13-14	40	1	M	1940-60	ETS	
English Expression, Part II Mechanics of Expression	13-14	40	1	M			
Reading Comprehension	13-14	40	1	M			

Table V (Continued)
Achievement: Language and Literature

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Cooperative Inter-American Tests of Language Usage</i> (for use with students of English or Spanish as a second language)							
English	8-13	35	2	M	1963	GTA	
Spanish	8-13	35	2	M			
<i>Greene-Stapp Language Abilities Test</i>	9-13	120	2	M	1952-54	HBW	5-195
LITERATURE							
<i>Center-Durost Literature Acquaintance Test</i>	11-13	40	1	M	1953	HBW	5-210
<i>Cooperative Literary Comprehension and Appreciation Test</i>	10-16	40	2	M	1935-51	ETS	4-184 3-142

Table VI
Achievement: Mathematics

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
ARITHMETIC AND GENERAL MATHEMATICS							
<i>California Arithmetic Test</i>	(See Achievement Test Batteries)						5-468
<i>Basic Skills in Arithmetic Test—</i> W. L. Wrinkle and others	6-12	40-45	2	H	1945	SRA	3-335
<i>A Brief Survey of Arithmetic Skills,</i> Revised Edition—A. E. Traxler	7-12	20	2	H	1947-53	BM	5-467
<i>Buswell-John Diagnostic Chart for</i> <i>Fundamental Processes in Arithmetic</i>	2-8	20	1	H	1925	BM	4-413 40-1456
<i>Cooperative General Achievement Tests:</i> <i>Test 3, Mathematics</i>	(See Achievement Test Batteries)						(5-420) (4-379) 3-316
<i>Cooperative General Mathematics Test for</i> <i>High School Classes</i>	11-13	40	1	M	1933-51	ETS	40-1432
<i>Cooperative Mathematics Test for Grades 7,</i> <i>8, and 9—B. Orshansky and H. V. Price</i>	7-9	80	2	M	1950	ETS	5-421 4-370
<i>Davis Test of Functional Competence</i> <i>in Mathematics</i>	9-13	80	2	M	1951-52	HBW	5-422
<i>Madden-Peak Arithmetic Computation Test</i>	7-12	49	2	M	1954-57	HBW	5-478
<i>New Cooperative Tests in Arithmetic</i>	7-9	40	3	M	1962	ETS	

Table VI (Continued)
Achievement: Mathematics

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>New York Test of Arithmetical Meanings—</i>							
<i>J. W. Wrightstone and others</i>							
Level One	1.9-2.1	60	1	H	1956	HBW	5-480
Level Two	2.9-3.1	60	1	H			
<i>Number Fact Check Sheet—R. Cochrane</i>	5-8	25	1	M	1946-47	CTB	4-417
<i>Snader General Mathematics Test</i>	9-13	40	2	M	1951-54	HBW	(5-439) 4-378
ALGEBRA							
<i>Blyth Second-Year Algebra Test</i>	10-12	45	2	M	1953-54	HBW	5-443
<i>Cooperative Algebra Test: Elementary</i>							
<i>Algebra Through Quadratics—</i>							
<i>M. M. Martin and others</i>	9-12	40	3	M	1932-51	ETS	4-387
<i>Cooperative Intermediate Algebra Test:</i>							
<i>Quadratics and Beyond</i>	9-12	40	3	M	1933-51	ETS	4-388
<i>Lankton First-Year Algebra Test</i>							
<i>(End-of-Year Test)</i>	9-12	40	2	M	1951-54	HBW	5-451 4-394
<i>Larson-Greene Unit Tests in</i>							
<i>First-Year Algebra</i>	9-12	8-20	2	H	1947	IOWA	4-395
<i>New Cooperative Tests in Algebra</i>							
Algebra I (Elementary)	9-12	40	2	M	1962	ETS	
Algebra II (Intermediate)	10-12	40	2	M	1962		
Algebra III (Advanced)	11-12	40	2	M	1964		

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Seattle Algebra Test (First Semester)</i> — H. B. Jeffery and others	9-12	40	2	M	1951-54	HBW	5-452
GEOMETRY							
<i>Cooperative Plane Geometry Test</i> — M. P. Martin and others	10-11	40	3	M	1932-51	ETS	4-423 3-357
<i>New Cooperative Test in Geometry</i>	10-12	80	2	M	1962	ETS	
<i>Seattle Plane Geometry Test (First Semester)</i> —H. B. Jeffery and others	10-12	45	2	M	1951-54	HBW	5-497
<i>Shaycoft Plane Geometry Test</i>	11-12	40	2	M	1951-54	HBW	(5-498) 4-433
MISCELLANEOUS MATHEMATICS							
<i>Cooperative Plane Trigonometry Test</i> — J. A. Long and others	11-14	40	2	M	1932-51	ETS	(4-438) 40-1474
<i>New Cooperative Test in Analytical Geometry</i>	12-14	40	2	M	1964	ETS	
<i>New Cooperative Test in Calculus</i>	12-14	40	2	M	1964	ETS	
<i>New Cooperative Test in Trigonometry</i>	11-14	40	2	M	1964	ETS	

Table VII
Achievement: Music

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Diagnostic Tests of Achievement in Music—</i> M. L. Kotick and T. L. Torgerson	4-12	60	2	M	1950	CTB	4-226
<i>The Farnum Music Notation Test</i>	7-9	10	1	H	1953	PSYCH	5-246
<i>Kwalwasser-Ruch Test of Musical Accomplishment</i>	4-12	50	1	H	1924-27	IOWA	40-1333
<i>Kwalwasser Test of Music Information and Appreciation</i>	7-16	40-45	1	H	1927	IOWA	40-1334

Table VIII
Achievement: Reading and Vocabulary

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Cooperative English Test:</i> <i>Reading Comprehension, 1960 Revision—</i>							
F. B. Davis and others							
Lower Level	9-12	40	2	M	1940-60	ETS	(5-645) (4-547) 3-497
Higher Level	13-14	40	2	M			
<i>Cooperative Inter-American Tests</i> <i>of Reading^a</i>							
Primary	1, 2, L-3	40	2	H	1959-64	GTA	
Intermediate	4-7	40	2	M			
Advanced	8-13	40	2	M			
<i>Cooperative Vocabulary Test—</i>							
F. B. Davis and others	7-16	30	2	M	1940-53	ETS	(4-213) 3-160
<i>Davis Reading Test</i>							
Series 1	11-13	40	4	M	1956-58	PSYCH	5-625
Series 2	8-11	40	4	M	1962		
<i>Diagnostic Reading Tests—</i>							
F. O. Triggs and others							
Survey Section						CDRT	
Booklet I	K-1		1	H			
Booklet II	2		2	H			
Booklet III	3-4		2	H			

Table VIII (Continued)
Achievement: Reading and Vocabulary

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
Survey Section							
Lower Level	4-8	48	4	M	1952-60	SRA	4-531
Upper Level	7-13	48	8	M	1947-60	SRA	4-531
Complete Battery							
Lower Level	4-8		2	M	1952-60	CDRT	4-531
Upper Level	7-13		2	M	1947-60	CDRT	4-531
<i>Doren Diagnostic Reading Test of Word Recognition Skills</i>	1-6	180	1	H	1956	AGS	5-659
<i>Durost-Center Word Mastery Test</i>	9-12	60	1	M	1950-52	HBW	5-233
<i>Durrell Analysis of Reading Difficulty, New Edition^b</i>	1-6	30-45	1	H	1937-55	HBW	5-660
<i>Durrell-Sullivan Reading Capacity and Achievement Tests</i>							
Primary Test	2.5-4.5	55-65	1	H	1937-45	HBW	5-661 4-562
Intermediate Capacity Test	3-6	30-40	1	H			
Intermediate Achievement Test	3-6	45-55	2	H			
<i>Gates Advanced Primary Reading Tests</i>							
Word Recognition	2-3	15	3	H	1926-58	TC-COL	(5-630) 3-484
Paragraph Reading	2-3	25	3	H			

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Gates Basic Reading Tests</i>							
GS—Reading to Appreciate General Significance	3.5-8	8-10	3	H	1958	TC-COL	5-631 3-485
UD—Reading to Understand Precise Directions	3.5-8	8-10	3	H			
ND—Reading to Note Details	3.5-8	8-10	3	H			
RV—Reading Vocabulary	3.5-8	20	3	H			
LC—Level of Comprehension	3.5-8	20	3	H			
<i>Gates Primary Reading Tests</i>							
Word Recognition	1-2.5	15	3	H	1926-58	TC-COL	(5-632) 4-563 3-486
Sentence Reading	1-2.5	15	3	H			
Paragraph Reading	1-2.5	20	3	H			
<i>Gates Reading Diagnostic Tests, Revised^b</i>	1-8	50	2	H	1926-53	TC-COL	5-662 4-563
<i>Gates Reading Survey</i>	3.5-10	45-60	3	H, M	1939-58	TC-COL	(5-633) 3-487
<i>Gilmore Oral Reading Test^b</i>	1-8	15-20	2	H	1951-52	HBW	5-671
<i>Gray Oral Reading Tests^b</i>	1-12		4	H	1915-63	BM	40-1571
<i>Iowa Silent Reading Tests: New Edition—</i>							
H. A. Greene and others	4-8	49	4	M	1933-56	HBW	3-489
Elementary	9-13	45	4	M	1927-43		
Advanced							

Table VIII (Continued)
Achievement: Reading and Vocabulary

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Kelley-Greene Reading Comprehension Test</i>	9-13	65-75	2	M	1953-55	HBW	5-636
<i>Michigan Vocabulary Profile Test</i>	9-16 Adults	50	2	M	1937-49	HBW	4-216 3-166 40-1320 38-1171
<i>Nelson Reading Test</i> , 1962 Edition	3-9	30	2	M, Q	1962	HM	4-545 3-492
<i>SRA Reading Record</i> — G. T. Buswell and M. Buswell	6-12	40	1	Q	1947-59	SRA	4-550 3-502

^a Available in English and Spanish editions. A new Inter-American series of tests of reading is being completed, with 1964 as the probable release date.

^b Individually administered.

Table IX
Achievement: Science

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
BIOLOGICAL SCIENCES							
<i>California Tests in Social and Related Sciences, Part III—G. S. Adams and others Advanced</i>	9-12	80	2	M, Q	1954-55	CTB	5-4
<i>Cooperative Biology Test—P. E. Kambly</i>	10-12	40	2	M	1933-51	ETS	4-601
<i>Cooperative Inter-American Tests of Natural Sciences (English and Spanish Editions)</i>	8-13	40	2	M	1959-64	GTA	
<i>Nelson Biology Test</i>	9-13	40	2	M	1951-54	HBW	5-728
<i>New Cooperative Test in Biology</i>	9-12	40	2	M	1964	ETS	4-605
<i>Survey Test in Biological Science—G. S. Adams and others</i>	7-10	40	1	M	1959	CTB	
GENERAL SCIENCE AND ELEMENTARY SCIENCE							
<i>Cooperative General Science Test—P. E. Kambly and C. A. Pearson</i>	9	40	3	M	1932-51	ETS	4-623
<i>Cooperative Science Test for Grades 7, 8, and 9—P. E. Kambly</i>	7-9	80	2	M	1941-51	ETS	4-624 3-571
<i>New Cooperative Test in General Science</i>	8-12	40	2	M	1964	ETS	
<i>New Cooperative Test in Advanced General Science</i>	9-12	40	2	M	1964	ETS	

Table IX (Continued)
Achievement: Science

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Read General Science Test</i>	9-10	40	2	M	1951-54	HBW	(5-715) 4-628
<i>Survey Test in Physical Science— G. S. Adams and others</i>	7-10	40	1	M	1959	CTB	
PHYSICS AND CHEMISTRY							
<i>Anderson Chemistry Test</i>	10-13	40	2	M	1951-54	HBW	5-737 4-613
<i>Cooperative Chemistry Test— P. J. Burke and J. F. Castka</i>	10-12	40	3	M	1933-50	ETS	5-744
<i>Cooperative Physics Test— P. J. Burke and others</i>	10-12	40	2	M	1932-51	ETS	5-751
<i>Dunning Physics Test</i>	10-13	45	2	M	1951-54	HBW	5-753 4-636
<i>New Cooperative Test in Chemistry</i>	11-12	40	2	M	1964	ETS	
<i>New Cooperative Test in Physics</i>	11-12	40	2	M	1964	ETS	
<i>Physical Science Study Committee Tests^a</i>	10-12	45	1	M	1959	ETS	
<i>Toledo Chemistry Placement Examination— N. W. Hovey and others</i>	13	55	1	M, Q	1959-63	TOLEDO	

^a Ten tests usable only by schools using the PSSC textbook.

Table X
Achievement: Social Studies

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>California Tests in Social and Related Sciences, Parts I and II—</i>							
G. S. Adams and others							
Elementary	4-8	120	2	M, Q	1946-55	CTB	5-4
Advanced	9-12	90	2	M, Q	1954-55		4-23 5-4
<i>Cooperative American Government Test—</i>							
J. Haefner	10-12	40	2	M	1947-51	ETS	4-702
<i>Cooperative American History Test—</i>							
H. D. Berg	11-14	40	3	M	1932-51	ETS	4-684
<i>Cooperative Inter-American Tests of Social Studies</i> (English and Spanish Editions)	8-13	40	2	M	1959-64	GTA	
<i>Cooperative Modern European History Test</i> —F. H. Stutz	10-14	40	2	M	1932-51	ETS	(4-686) 40-1635 38-1016
<i>Cooperative Social Studies Test for</i> <i>Grades 7, 8, and 9—</i> H. D. Berg and others	7-9	80	2	M	1941-51	ETS	4-663
<i>Cooperative World History Test—</i> W. Taylor and F. H. Stutz	10-11	40	2	M	1934-51	ETS	5-814
<i>Crary American History Test</i>	9-13	40	2	M	1950-54	HBW	5-816 4-688

Table X (Continued)
Achievement: Social Studies

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Cummings World History Test</i>	9-13	40	2	M	1950-54	HBW	(5-817) 4-689
<i>Dimond-Pflieder Problems of Democracy Test</i>	9-13	40	2	M	1952-54	HBW	5-833
<i>Engle Psychology Test</i>	11-13	40	2	M	1952-54	HBW	5-582
<i>Peltier-Durost Civics and Citizenship Test</i>	9-12	55	2	M	1958	HBW	(5-840)

Table XI
Aptitude: Group Tests of General Mental Ability

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Academic Promise Tests (APT)</i> — G. K. Bennett and others	6-9	90	2	M	1962	PSYCH	
<i>California Short Form Test of Mental Maturity^a</i> (CTMM-S)—E. T. Sullivan and others	K-1	20	1	Q	1938-58	CTB	5-313 4-282
Preprimary							
Primary	1-3	20	1	Q			
Elementary	4-8	20	1	M, Q			
Junior High School	7-9	20	1	M, Q			
Secondary	9-13	53	2 ^b	M, Q			
Advanced	10-16	53	1	M, Q			
	Adults						
<i>California Test of Mental Maturity^c</i> — 1957 Edition (CTMM)							
Preprimary	K-1		1	Q	1956-57	CTB	5-314 4-282
Primary	1-3		1	Q			
Elementary	4-8		1	M, Q			
Secondary	9-13		1	M, Q			
Advanced	10-16		1	M, Q			
	Adults						
<i>Chicago Non-Verbal Examination</i>	Age 6-over ^d	25	1	H	1936-54	PSYCH	5-316 40-1387
<i>College Qualification Tests^e</i> — G. K. Bennett and others	11-13	80	3	M	1955-60	PSYCH	5-320

Table XI (Continued)
Aptitude: Group Tests of General Mental Ability

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Cooperative Inter-American Tests of General Ability^c</i>							
Primary	1, 2, L3	35	2	H	1959-64	GTA	
Advanced	4-7	35	2	M			
Intermediate	8-13	35	2	M			
<i>Cooperative School and College Ability Tests^{ac} (SCAT)</i>							
Level 5	4-6	70	2	M	1955-57	ETS	5-322
Level 4	6-8	70	2	M			
Level 3	8-10	70	2	M			
Level 2	10-12	70	2	M			
Level 1	12-14	70	4	M			
<i>Davis-Eells Test of General Intelligence or Problem-Solving Ability</i>							
Primary	1-2	60	1	H	1953	HBW	5-326
Elementary	3-6	90	1	H			
<i>Harris-Goojenough Test of Psychological Maturity</i>							
	K-3	5-10	1	H	1926-61	PSYCH	(5-335) 4-292
<i>Henmon-Nelson Tests of Mental Ability, Revised Edition</i>							
Grades 3-6	3-6	30	2	M, Q	1957	HM	5-342
Grades 6-9	6-9	30	2	Q			
Grades 9-12	9-12	30	2	Q			
Grades 13-17	13-17	30	2	Q			

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Kuhlmann-Anderson Intelligence Tests, Sixth Edition</i>							
K	K	20-30	1	H	1927-52	PP BM	5-348 4-302
A	1	20-30	1	H			
B	2	20-30	1	H			
C	3	20-30	1	H			
D	4	20-30	1	H			
E	5	20-30	1	H			
F	6	20-30	1	H			
G	7-8	20-30	1	H			
<i>Kuhlmann-Anderson Intelligence Tests, Seventh Edition</i>							
D	4-5	45	1	M	1927-60	PP BM	
EF	5-7	45	1	M			
G	7-9	45	1	M			
H	9-12	45	1	M			
<i>The Lorge-Thorndike Intelligence Tests¹</i>							
Level 1	K-1	20	2	H	1949-59	HM	5-350
Level 2	2-3	20	2	H			
Level 3	4-6	} Verbal-34 Nonverbal-27	2	M			
Level 4	7-9		2	M			
Level 5	10-12		2	M			
<i>Ohio State University Psychological Test, Form 21¹—H. A. Toops</i>							
	9-16 Adults	120	2	M, Q	1919-59	SRA	5-359 4-308

Table XI (Continued)
Aptitude: Group Tests of General Mental Ability

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Otis Quick-Scoring Mental Ability Tests,</i> New Edition							
Alpha	1.5-4	40	2	Q	1936-54	HBW	5-362
Alpha-Short Form	1.5-4	25	1	Q			
Beta	4-9	30	4	M, Q			
Gamma	9-16	30	4	M, Q			
<i>Pintner-Cunningham Primary</i>	K-2	25	3	H	1923-46	HBW	(5-368) 3-255 40-1416
<i>Pintner-Durost Elementary Test</i>							
Scale 1 Picture Content	2-4	45	2	Q	1940-41	HBW	(5-368)
Scale 2 Reading Content	2-4	45	2	Q			3-255
<i>Pintner General Ability Tests:</i>							
<i>Nonlanguage Series</i>	4-9	50-60	2		1945	HBW	3-254
<i>Pintner General Ability Test:</i>							
<i>Verbal Series</i>							
Intermediate	5-9	45	2	M	1938-42	HBW	5-329
Advanced	9-13	55	2	M			
<i>Raven Progressive Matrices—Form 1938</i>	Ages 8-Adults	45	1	H	1938-58	PSYCH	(5-370) 4-314 3-258 40-1417
Form 1947 (colored matrices)	Ages 5-11 and Defective Adults	15-30	1	H	1947-56		

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Revised Beta Examination</i> ^j — D. E. Kellogg and N. W. Morton	Ages 16 and over	30	1	H	1931-57	PSYCH	(5-375) 3-259 40-1419
<i>SRA Non-Verbal Form</i> — R. N. McMurry and J. E. King	Ages 12 and over	15-20	1	Q	1946-47	SRA	4-318
<i>SRA Primary Mental Abilities, Revised</i> ^k — L. L. Thurstone and T. G. Thurstone							
Grades K-2	K-2	50-60	1	H	1946-62	SRA	5-614 4-716
Grades 2-4	2-4	50-60	1	H			
Grades 4-6	4-6	50-60	1	M			
Grades 6-9	6-9	45-55	1	M			
Grades 9-12	9-12	45-55	1	M			
<i>SRA Tests of Educational Ability (TEA)</i> — L. L. Thurstone and T. G. Thurstone ^l							
	4-6	60	1	M	1957-58	SRA	5-377
	6-9	65	1	M			
	9-12	45-50	1	M			
<i>SRA Tests of General Ability (TOGA)</i> ^m							
K-2	K-2	35-45	1	H	1957-60	SRA	
2-4	2-4	35-45	1	H			
4-6	4-6	35-45	1	M			
6-9	6-9	35-45	1	M			
9-12	9-12	35-45	1	M			
<i>SRA Verbal Form</i>	Ages 12 and over	15-20	2	Q	1946-56	SRA	(5-378) 4-319

Table XI (Continued)
Aptitude: Group Tests of General Mental Ability

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Survey of Mental Maturity</i> ⁿ — W. W. Clark and others							
Jr. High Level	7-9	30	2	M	1959	CTB	
Advanced	10-12 Adults	30	2	M			
<i>Terman-McNemar Test of Mental Ability</i>	7-12	40	1	M	1941-42	HBW	(4-324) 3-263

^a Seven scores: spatial relationships, logical reasoning, numerical reasoning, verbal concepts, language, nonlanguage, total.

^b A restricted form (B) is available for use in scholarship testing or other special programs.

^c Eight scores: memory, spatial relationships, logical reasoning, numerical reasoning, verbal concepts, language, nonlanguage, total.

^d A group intelligence test designed for children handicapped in their use of the English language (the deaf, those with reading difficulties, and the like). Verbal directions usable for age 6 to adult, pantomime directions for age 8 to adult.

^e Verbal, numerical, and information subtests available as separates. Information test has three scores: science, social science, total.

^f Available in English and Spanish editions. A new *Inter-American Series of Tests of General Ability* is being completed, with 1964 as a probable release date.

^g Three scores: verbal, quantitative, total.

^h Levels 3, 4, and 5 have verbal and nonverbal subtests. Spanish directions are available for nonverbal battery.

ⁱ Four scores: same-opposites, analogies, reading comprehension, total.

^j French edition available.

^k Batteries for two lower levels include subtests on verbal meaning, spatial ability, perception, and quantitative ability. The battery for grades 4-6 includes these plus reasoning. The batteries for the two highest levels include verbal meaning, number sense, reasoning, and space.

^l Four scores: language, reasoning, quantitative, total.

^m Spanish edition available.

ⁿ Three scores: language, nonlanguage, total.

Table XII
Aptitude: Individual Tests of General Mental Ability
 (To be used only by qualified psychologists or by specially trained psychometrists
 working under their supervision)

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCOR- ING	PUBLICATION DATLS	PUBLISHER	REVIEWS
<i>Arthur Point Scale of Performance Tests</i> , Revised, Form II	Ages 5 and over		1	H	1933-47	PSYCH	4-335 40-1379
<i>Columbia Mental Maturity Scale, Revised</i> ^a —B. B. Burgemeister and others	MA 3-10 years	15-20	1	H	1954-59	HBW	(5-402)
<i>Leiter International Performance Scale</i>	Ages 2-18		1	H	1936-52	STOELTING	(5-408) 4-349
<i>Leiter International Performance Scale:</i> <i>Arthur Adaptation</i>	Ages 2-12		1	H	1952-55	STOELTING	(5-407)
<i>Nebraska Test of Learning Aptitude</i> — M. S. Hiskey ^b	Ages 4-10		1	H	1941-55	AUTHOR ^b	5-409
<i>Peabody Picture Vocabulary Test</i> — L. M. Dunn	Ages 2½-18	15	2	Q	1959	AGS	
<i>Revised Stanford-Binet Scales</i> , Third Edition, 1960—L. M. Terman and M. Merrill	Ages 2 and over		1	H	1916-60	HM	5-413 4-358
<i>Wechsler Adult Intelligence Scale (WAIS)</i>	Ages 16 and over		1	H	1939-55	PSYCH	5-414

Table XII (Continued)
 Aptitude: Individual Tests of General Mental Ability
 (To be used only by qualified psychologists or by specially trained psychometrists
 working under their supervision)

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCOR- ING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Wechsler Intelligence Scale for Children</i> (WISC) ^c	Ages 5-15		1	H	1949	PSYCH	5-416 4-363

^a Usable with children with severe motor handicaps.

^b Norms available for deaf and hard-of-hearing children attending residential state schools for the deaf. Published by the author (Marshall S. Hiskey, 5640 Baldwin, Lincoln, Neb.).

^c Spanish edition of manual and record forms available.

Table XIII
Aptitude Test Batteries

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Differential Aptitude Tests (DAT)—</i>							
G. K. Bennett and others ^a							
Verbal Reasoning	8-13	30	2	M	1947-63	PSYCH	5-605 4-711
Numerical Ability	8-13	30	2	M			
Abstract Reasoning	8-13	25	2	M			
Space Relations	8-13	30	2	M			
Mechanical Reasoning	8-13	30	2	M			
Clerical Speed and Accuracy	8-13	6	2	M			
Language Usage	8-13	35	2	M			
<i>Flanagan Aptitude Classification Tests (FACT)^b</i>							
	9-12 Adults	3 half-day sessions	1	Q	1951-60	SRA	5-608
<i>General Aptitude Test Battery (GATB)^c</i>							
	9-12 Adults	120-150 (for group tests)	1	M, Q	1946-59	U.S. Empl. Service	5-609
<i>Guilford-Zimmerman Aptitude Survey</i>							
I. Verbal Comprehension	9-16 Adults	25	1	M	1956	SHERIDAN	4-715
II. General Reasoning	9-16 Adults	35	1	M			
III. Numerical Operations	9-16 Adults	8	1	Q			
IV. Perceptual Speed	9-16 Adults	5	1	Q			

Table XIII (Continued)
Aptitude Test Batteries

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
V. Spatial Orientation	9-16 Adults	10	1	M			
VI. Spatial Visualization	9-16 Adults	30	1	M			
VII. Mechanical Knowledge	9-16 Adults	50	1	M			
<i>Holzinger-Crowder Uni-Factor Tests</i> ^d	7-12	80-90	1	M	1952-55	HBW	5-610
<i>Multiple Aptitude Tests (MAT)</i> , 1959 Edition—D. Segel and E. Raskin ^e	7-13	175-220	1	M	1955-60	CTB	5-613

^a Spanish edition available.

^b Available as separates: Inspection, Coding, Memory, Precision, Assembly, Scales, Coordination, Judgment and Comprehension, Arithmetic, Patterns, Components, Tables, Mechanics, Expression, Reasoning, Ingenuity.

^c Available only through the state employment service. Subtests on: Name Comparison, Computation, Three-Dimensional Space, Vocabulary,

Tool Matching, Arithmetic Reasoning, Form Matching, Mark Making, Pegboard (placing and turning), Finger Dexterity Board (assembling, disassembling).

^d Five scores: verbal, spatial, numerical, reasoning, scholastic aptitude.

^e Three scores in verbal comprehension, three scores in perceptual speed, three scores in numerical reasoning, four scores in spatial visualization.

Table XIV
Aptitude: Art

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Graves Design Judgment Test</i>	7-16 Adults	20-30	1	M	1948	PSYCH	4-220
<i>Horn Art Aptitude Inventory</i>	12-16 Adults	50	1	H	1939-53	STOELTING	5-242 3-171
<i>Meier Art Tests</i>							
I—Art Judgment	7-16 Adults	40-60	1	M	1929-42	IOWA	4-224 3-172
II—Aesthetic Perception	7-16 Adults	40-60	1	Q	1963-64		
<i>Tests in Fundamental Abilities of Visual Art</i> —A. S. Lewerenz	3-12 Adults	85	1	H	1927	CTB	40-1329

Table XV
Aptitude: Business-Clerical

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>E. R. C. Stenographic Aptitude Test—</i> W. H. Deemer	10-12 Adults	33	1	H	1944	SRA	3-372
<i>Minnesota Clerical Test</i> ^a — D. M. Andrew and others	8-12 Adults	15	1	H	1933-59	PSYCH	5-850 3-627 40-1664
<i>Psychological Corporation General Clerical Test</i>	9-16 Adults	43	1	H	1944-50	PSYCH	4-730 3-630
<i>SRA Clerical Aptitudes</i>	9-12 Adults	35	1	Q	1947-50	SRA	4-732
<i>Stenographic Aptitude Test—</i> G. K. Bennett	9-16	25	1	H	1939-46	PSYCH	3-390
<i>Survey of Working Speed and Accuracy—</i> F. Ruch	9-16 Adults	20	1	H	1943-48	CTB	3-631
<i>Turse Shorthand Aptitude Test</i>	8-12 Adults	40	1	H	1937-40	HBW	4-460 3-393
<i>Turse Clerical Aptitudes Test</i>	8-12 Adults	28	1	H	1955	HBW	5-855

^a A Spanish adaptation is available.

Table XVI
Aptitude: Foreign Language

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Foreign Language Prognosis Test—</i> M. Symonds	8-9	44	2	H	1930-59	TC-COL	4-232 40-1340
<i>Iowa Placement Examinations:</i> <i>Foreign Language Aptitude Series FA-2,</i> Revised—G. D. Stoddard and others	12-13	45	3	M	1925-44	IOWA	3-178
<i>Modern Language Aptitude Test—</i> J. B. Carroll and S. M. Sapon	9-16 Adults	30-60 ^a	1	M	1958-59	PSYCH	

^a Thirty minutes' working time for Short Form, 60 minutes for total test.

Table XVII
Aptitude: Manual Dexterity and Mechanical Aptitude

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Bennett Hand-Tool Dexterity Test</i> ^a	H.S. Adults	4-12	1	H	1946	PSYCH	3-659
<i>Crawford Small Parts Dexterity Test</i> ^a	H.S. Adults	9-25	1	H	1946-56	PSYCH	5-871 4-752 3-667
<i>Minnesota Rate of Manipulation Test</i>	H.S. Adults	10-15	1	H	1931-57	AGS	3-663 40-1662
<i>O'Connor Finger Dexterity Test</i> ^a	11-12 Adults	8-15	1	H	1920-26	STOELTING	40-1659
<i>O'Connor Tweezer Dexterity Test</i> ^a	11-12 Adults	6-10	1	H	1920-28	STOELTING	40-1678
<i>O'Rourke Mechanical Aptitude Test</i>	H.S. Adults	55	2	H	1926-57	PSYCH	(5-882) 3-672
<i>Prognostic Test of Mechanical Abilities—</i> J. W. Wrightstone and C. E. O'Toole	7-12 Adults	38	1	M	1946-47	CTB	4-761
<i>Purdue Pegboard</i> ^a —J. Tiffin	9-16 Adults	10	1	H	1941-48	SRA	5-873 4-751 3-666

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Revised Minnesota Paper Form Board</i> — R. Likert and W. H. Quasha ^b	9-16 Adults	20	2	M	1930-48	PSYCH	5-884 4-763 3-677 40-1673
<i>SRA Mechanical Aptitudes</i>	9-12 Adults	40	1	Q	1950	SRA	4-764
<i>Stromberg Dexterity Test</i> ^a	H.S. Adults	5-10	1	H	1945-51	PSYCH	4-755
<i>Mechanical Comprehension Tests</i> ^c — G. K. Bennett and others							
Form AA	H.S. boys Adults	25-35	1	M	1940-55	PSYCH	(5-889) 4-766 3-683
Form BB (more difficult)	Applicants for Technical Training or Employment	25-35	1	M	1941-51		
Form CC (most difficult)	Engineering Students	25-35	1	M	1949		
Form W-1	H.S. girls Adults	25-35	1	M	1942-47		

^a Individually administered.

^b A bilingual edition (with instructions in both French and English) has been prepared for French-Canadian use. A form with Spanish instructions is also available.

^c Spanish editions of Forms AA and BB available; bilingual English-French edition of Form AA available for French-Canadian use.

Table XVIII
Aptitude: Mathematics

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>California Algebra Aptitude Test—</i> N. Keys and M. McCrum	8-12	50	1	H	1940-58	AGS	(5-444) 4-385 3-320
<i>Orleans Algebra Prognosis Test,</i> Revised Edition	7-9	39	1	H	1928-51	HBW	4-396 40-1444
<i>Orleans Geometry Prognosis Test,</i> Revised Edition	9-11	39	1	H	1929-51	HBW	4-427 40-1471
<i>Survey Test of Algebraic Aptitude—</i> R. E. Dinkel	7-9	40	1	M	1959	CTB	

Table XIX
Aptitude: Music

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Drake Musical Aptitude Tests</i>	3-16 Adults	45-60	2	Q	1954-57	SRA	5-245 3-175
<i>Musical Aptitude Test—</i> H. S. Whistler and L. P. Thorpe	4-10	40	1	M	1950	CTB	5-250 4-228
<i>Seashore Measures of Musical Talent,</i> Revised Edition	4-16 Adults	60-70	1	M	1919-60	PSYCH	5-251 4-229 3-177 40-1338
<i>Wing Standardized Tests of</i> <i>Musical Intelligence</i>	5-16 Adults	50-70	1	H	1939-60	Natl. Found. ^a	5-254

^a Distributed by the National Foundation for Educational Research in England and Wales. (79 Wimpole St., London W.1., England).

Table XX
Aptitude: Reading Readiness

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Gates Reading Readiness Tests</i>	K-1	50	1	H	1939-42	TC-COL	4-566 3-516
<i>Harrison-Stroud Reading Readiness Profiles</i>	K-1	79	1	H	1949-56	HM	5-677
<i>Metropolitan Readiness Tests— G. Hildreth and N. L. Griffiths</i>	K-1	65-75	2	H	1933-50	HBW	4-570
<i>Monroe Reading Aptitude Tests</i>	K-1	30-40	1	H	1935	HM	3-519
<i>Murphy-Durrell Diagnostic Reading Readiness Test</i>	K-1	80	1	H	1947-49	HBW	5-679

Table XXI
Aptitude: Science

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Engineering and Physical Science Aptitude Test</i> —B. V. Moore and others	12-16 Adults	72	1	M	1951	PSYCH	4-810 3-698
<i>Physical Scientific Aptitude Examination: Form S</i> —J. Lapp and others	12-13	60	1	H	1943	IOWA	3-547
<i>The Purdue Physical Science Aptitude Test</i> —H. H. Remmers and N. A. Rosen	9-13	60	2	M	1943-60	PURDUE	

Table XXII
Interest Inventories (to be used only under the supervision of qualified psychologists
or counselors with appropriate training)

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Kuder Preference Record— Occupational-Form D</i>	9-16 Adults	20-30	1	M ^a	1956-59	SRA	5-862
<i>Kuder Preference Record— Vocational-Form C</i>	9-16 Adults	40-45	2	M, Q	1934-56	SRA	5-863 4-742
<i>Occupational Interest Inventory (OII)— E. A. Lee and L. P. Thorpe</i>	7-16 Adults	30-40	1	M, Q	1943-56	CTB	5-864
Intermediate							
Advanced	9-16 Adults	30-40	1	M, Q			
<i>Picture Interest Inventory—K. P. Weingarten</i>	7-12 Adults	30-40	1	M	1958	CTB	(5-865)
<i>A Study of Values— G. W. Allport and others</i>	13-16 Adults	20-30	1	H	1931-60	HM	5-114 4-92
<i>Vocational Interest Analyses— E. C. Roeber and others</i>	9-12 Adults	45	1	M		CTB	5-870 4-746

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Vocational Interest Blank for Men,</i> Revised Form M—E. K. Strong (SVIB)	Ages 17 and over	30–60	1	M	1927–59	CPP	5-868 4-747 3-647 40-1680
<i>Vocational Interest Blank for Women,</i> Revised Form W—E. K. Strong	Ages 17 and over	30–60	1	M	1933–59	CPP	5-869 3-649

^a Keys available for 48 predominantly masculine occupations.

Table XXIII
 Personal-Social Adjustment (to be used only under the supervision of qualified psychologists
 or counselors with appropriate training)

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>The Adjustment Inventory</i> —H. M. Bell Student Form	9-16 Adults	20-30	1	M, Q	1934-39	CPP	(5-30) 4-28 40-1200 38-912
Adult Form	9-16 Adults	20-30	1	M, Q	1938-39	CPP	
<i>Bell Adjustment Inventory</i> , Revised Student Form	9-16	30-40	1	M	1958	CPP	
<i>Bernreuter Personality Inventory</i>	9-16 Adults	20-30	1	M	1931-38	CPP	5-95 4-77 40-1239
<i>Billett-Starr Youth Problems Inventory</i> Junior Level	7-9	60-75	1	Q	1961	HBW	
Senior Level	10-12	60-75	1	Q			
<i>California Psychological Inventory (CPI)</i> — H. Gough	8-16	45-60	1	M, Q	1956-57	CPP	5-37
<i>California Test of Personality (CTP)</i> — L. P. Thorpe and others Primary	K-3	40-50	2	H	1953	CTB	5-38 3-26
Elementary	4-8	40-50	2	M, Q			
Intermediate	7-10	40-50	2	M, Q			
Secondary	9-16	40-50	2	M, Q			

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Edwards Personal Preference Schedule</i>	13-16 Adults	45	1	M, Q	1953-59	PSYCH	5-47
<i>Gordon Personal Inventory</i>	9-16 Adults	15	1	H	1956	HBW	5-58
<i>Gordon Personal Profile</i>	9-16 Adults	15	1	H	1953-54	HBW	5-59
<i>Guilford-Zimmerman Temperament Survey</i>	9-16 Adults	45	1	M	1949-55	SHERIDAN	5-65 4-49
<i>Kuder Preference Record, Personal— Form A</i>	9-16 Adults	40-50	2	M, Q	1948-54	SRA	5-80 4-65
<i>Mental Health Analysis— L. P. Thorpe and W. C. Clark</i>							
Elementary	4-8	45-50	1	M	1946-59	CTB	3-59
Intermediate	7-10	45-50	1	M			
Secondary	9-16	45-50	1	M			
<i>Minnesota Counseling Inventory— R. F. Berdie and W. L. Layton</i>	8-12	45-50	1	M	1953-57	PSYCH	(5-85)
<i>Minnesota Multiphasic Personality Inventory (MMPI)^a—S. R. Hathaway and J. C. McKinley</i>	11-16 Adults	30-90	2 ^b	M	1943-51	PSYCH	5-86 4-71 3-60
<i>Mooney Problem Check List, 1950 Rev.^c— R. L. Mooney and L. V. Gordon</i>							
Form J	7-9	30-50	1	M	1942-50	PSYCH	(5-89) 4-73 3-67

Table XXIII (Continued)
 Personal-Social Adjustment (to be used only under the supervision of qualified psychologists
 or counselors with appropriate training)

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	DATES	PUBLISHER	REVIEWS
Form H	9-12	30-50	1	M	1941-50		
Form C	13-16	30-50	1	M	1941-50		
Form A	Adults	30-50	1	M	1950		
<i>SRA Junior Inventory—</i> H. H. Remmers and R. H. Bauernfeind	4-8	40	1	Q	1951-57	SRA	5-104 4-90
<i>SRA Survey of Interpersonal Values—</i> L. V. Gordon	9-16 Adults	15	1	H	1960	SRA	
<i>SRA Youth Inventory—</i> H. H. Remmers and others	7-12	40	2	M, Q	1949-56	SRA	(5-105) 4-91
<i>A Study of Values—</i> G. W. Allport and others	13-16 Adults	20	1	H	1931-60	HM	5-114 4-92
<i>Syracuse Scales of Social Relations—</i> E. F. Gardner and G. Thompson							
Grades 5-6	5-6	50-60	1	Q	1958-59	HBW	
Grades 7-9	7-9	50-60	1	Q			
Grades 10-12	10-12	50	1	Q			

^a Spanish edition of booklet form available.

^b An individually-administered form (printed on cards) and a group test (printed in a test booklet).

^c An adaptation for rural young people (ages 16-30) and an adaptation for student nurses may be purchased from Publications Office, Ohio State University.

Table XXIV
Work-Study Habits and Skills

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.	NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Brown-Holtzman Survey of Study Habits and Attitudes</i>	9-16	25-35	1	M	1953-56	PSYCH	5-688
<i>California Study Methods Survey— H. D. Carter</i>	7-13	35-50	1	M	1958	CTB	(5-689)
<i>Spitzer Study Skills Test</i>	9-13	150	2	M	1954-55	HBW	5-697
<i>Tyler-Kimber Study Skills Test</i>	9-16	60-90	1	H	1937	CPP	40-1580 38-1166

Table XXV
Miscellaneous

NAME OF TEST	GRADE LEVEL	WORKING TIME IN MIN.		NO. OF FORMS	SCORING	PUBLICATION DATES	PUBLISHER	REVIEWS
<i>Brown-Carlson Listening Comprehension Test</i>	9-12	45-50		2	M	1953-55	HBW	5-577
<i>Watson-Glaser Critical Thinking Appraisal</i>	9-16 Adults	30-45		1	M	1942-63	HBW	5-700

APPENDIX B

Publishers of Standardized Tests

AGS	American Guidance Service 720 Washington Ave., S.E. Minneapolis 14, Minn.	HM	Houghton Mifflin Co. 2 Park St. Boston 7, Mass.
BM	Bobbs-Merrill Co., Inc. 4300 W. 62 St. Indianapolis, Ind.	IND	Indiana State High School Testing Service Purdue University Lafayette, Ind.
CDRT	Committee on Diagnostic Reading Tests 419 West 119 St. New York 27, N. Y.	IOWA	Bureau of Educational Research and Service State University of Iowa Iowa City, Iowa
CPP	Consulting Psychologists Press, Inc. 557 College Ave. Palo Alto, Calif.	IPAT	Institute for Personality and Ability Testing 1602 Coronado Dr. Champaign, Ill.
CTB	California Test Bureau Del Monte Research Park Monterey, Calif.	KSTC	Bureau of Educational Measurements Kansas State Teachers College Emporia, Kans.
ETS	Educational Testing Service 20 Nassau St. Princeton, N. J.	MINN	University of Minnesota Press Minneapolis 14, Minn.
GTA	Guidance Testing Associates 6516 Shirley Ave. Austin 5, Texas	NBEA	National Business Education Assn. 1201 Sixteenth St., N.W. Washington 6, D. C.
HBW	Harcourt, Brace & World 757 Third Ave. New York 17, N. Y.	PA	Psychometric Affiliates Box 1625 Chicago 90, Ill.

PP	Personnel Press, Inc. 188 Nassau St. Princeton, N. J.	STOEL- TING	C. H. Stoelting Company 424 North Homan Ave. Chicago 24, Ill.
PSYCH	The Psychological Corporation 304 East 45 St. New York 17, N. Y.	TC-COL	Bureau of Publications Teachers College Columbia University New York 27, N. Y.
PURDUE	Purdue University Bookstore 360 State St. West Lafayette, Ind.	TOLEDO	The Research Foundation University of Toledo 2801 West Bancroft St. Toledo 8, Ohio
SHERI- DAN	Sheridan Supply Company P.O. Box 837 Beverly Hills, Calif.	USES	U.S. Employment Service Tests available to schools only when they are used in cooperation with State Employment Serv- ice offices.
SRA	Science Research Associates, Inc. 259 East Erie St. Chicago 11, Ill.	VET	Veterans' Testing Service American Council on Education 1785 Massachusetts Ave., N.W. Washington 6, D. C.
STAN	Stanford University Press Palo Alto, Calif.		

APPENDIX C

Methods of Expressing Test Scores (Based on the Normal Curve)

In Chapter 2, the reader was introduced to the normal curve and to the standard deviation (*SD*). The symbol σ (sigma) is used in Figure A.1 to represent the standard deviation. By using the standard deviation as a unit, it is possible to compare students' scores on tests of intelligence, various aspects of achievement, and other characteristics by ascertaining the position of each score or other datum in a normal frequency distribution.

In Chapter 2 the reader was also introduced to the concept of standard score, by means of which test scores are expressed in terms of their deviation from the average in standard-deviation units. It was explained that in order to avoid the negative numbers and decimal points of the original standard scores, several systems of equated scores have been developed, based on the use of 50, 100, or some other arbitrary number to represent the average, and 10, 20, or some other arbitrary number to represent the standard deviation. Figure A.1 illustrates the fundamental equivalence of various systems of equated scores.

In Chapter 6, the deviation IQ was described as a type of normalized standard score, by which a student's intelligence-test score could be compared with the scores of other members of his age group: if he were tested at the age of 10, with those of other 10-year-olds; if he were retested five years later, with those of other 15-year-olds. In each case, his raw score on the test would be transformed into a percentile rank within his own age group. Then his percentile rank within a normal distribution of 10-year-olds and 15-year-olds, respectively, would be translated into a type of normalized standard score (known, in this case, as a deviation IQ).

The standard-score equivalent (z) of a student's raw score is obtained by computing the deviation of that score from the mean and dividing that deviation by the standard deviation ($z = \frac{X - M}{SD}$). In other words, the z -score indicates the difference between a raw score and the mean of the distribution, expressed in standard-deviation units. A z score of +1 is one standard deviation above the mean; a z score of -1.5 is one and one-half standard deviations below the mean.

This same principle is used in setting up all systems of equated scores based

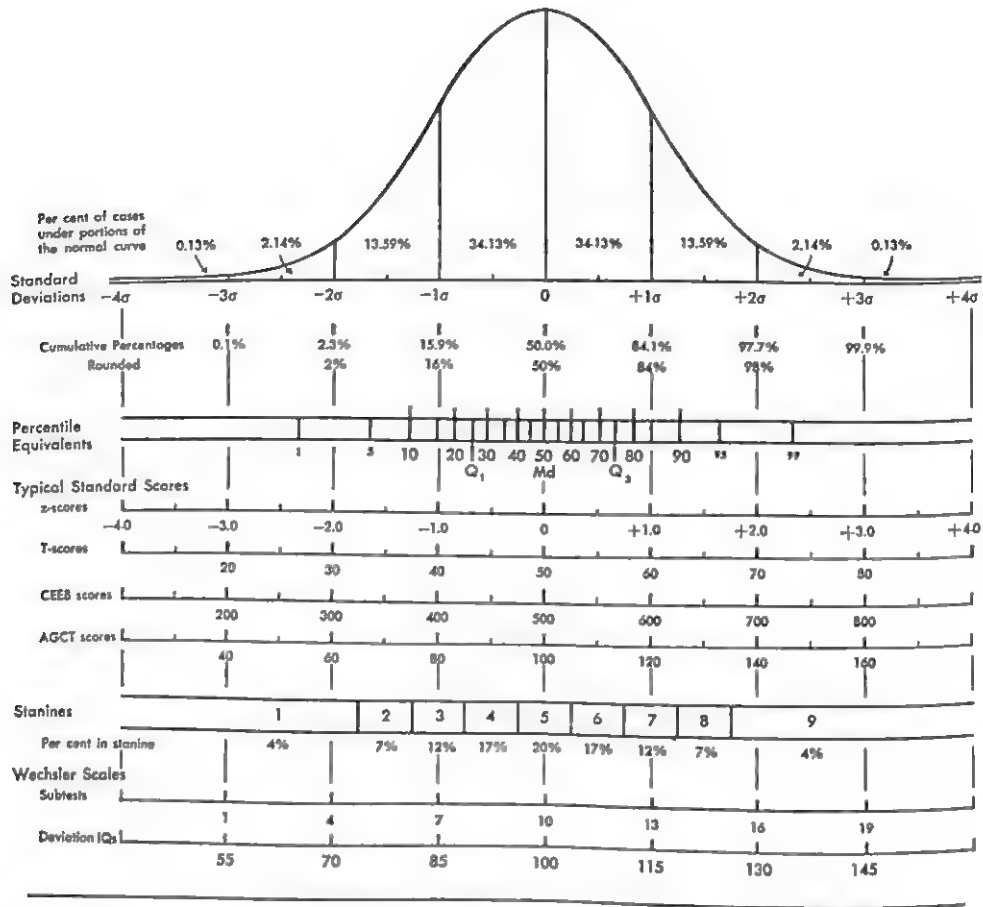


Fig. A.1 Chart Showing the Equivalences of Various Systems of Equated or Standard Scores.

Reproduced with the permission of the publisher from *Test Service Bulletin* No. 48 (New York: The Psychological Corporation, 1955).

on the normal curve—whether they are percentile ranks, *T*-scores, AGCT (Army General Classification Test) scores, or deviation IQ's. Examine the normal curve in Figure A.1. Note that there are no raw scores printed on the baseline. Hence, the person setting up a system of standard scores is free to use any numerical scale he chooses.¹ If he wishes to use regular standard scores (*z*-values), he sets the mean raw score of his test results equal to zero.

¹ However, the development of so many different kinds of standard scores is deplored in *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (Washington, D. C.: American Psychological Association, 1954). The committee recommends that the *T*-score (with a *M* of 50 and a *SD* of 10) be used when a two-digit standard score is desired; and that stanines (with a *M* of 5 and a *SD* of 2) be used when a one-digit standard score is desired.

Thus, if a distribution of scores on a specific test has a mean of 36 and a standard deviation of 4, one would enter on the baseline of the normal curve a raw score of 36 at the zero point or mean; a raw score of 40 ($36 + 4$) at a baseline position one unit to the right ($+1\sigma$); a raw score of 44 ($36 + 8$) at a position two units to the right ($+2\sigma$); a raw score of 32 ($36 - 4$) at a baseline position one unit to the left (-1σ); and the like. Once these entries have been made, one can read off intermediate standard-score values for all raw scores; for example, a raw score of 38 would be $+1.5$; a raw score of 41 would be 1.25 ; a raw score of 34, $-.5$; and the like.

In using the chart to determine the percentile ranks equivalent to certain raw scores, one makes use of the fact that *the total area under the curve represents the total number of cases* (in this case, test scores) in the frequency distribution. Vertical lines have been drawn through the score scale (the baseline) at the zero point and at points 1, 2, 3, and 4 standard-deviation units to the right (above the mean) and to the left (below the mean). These lines mark off subareas of the total area under the curve; that is, they mark off the score limits for certain percentages of cases in a normal frequency distribution. The numbers printed in these subareas indicate the percentages of students with scores falling within the specified limits. For example, 34.13 percent of all the students in a normally distributed group have scores falling between the mean (0) and $+1\sigma$. The fact that 68.26 percent (or approximately two thirds) of the cases fall between $+1\sigma$ and -1σ in a normal distribution was emphasized in Chapter 2.

Just below the standard-deviation scale on the chart is a row of percentages. This shows the cumulated percentages to the left or *below* each of the positions on the baseline. Thus, starting from the left, one sees that 0.1 percent of the individuals in a normal distribution have raw scores that would place them below -3σ ; 2.3 percent would be below -2σ ; 16 percent, below -1σ ; 50 percent, below 0, or the mean; 84 percent, below $+1\sigma$; and the like. The rounded values for these percentages are shown in the next row.

The next scale below the chart is for percentile ranks or percentile equivalents. If the lines for each of these percentile ranks were extended upward, through the area of the normal curve, one could visualize the principle, stressed on page 30, that percentile ranks near the average—for example, 55 and 65—represent almost identical raw scores, even though they are 10 percentile ranks apart; whereas a similar difference near either extreme of the distribution (for example, the difference between percentile ranks of 1 and 10 or of 90 and 99) represents a much larger difference in raw scores. One can readily see from the chart that 10 percent of the area (students) near the middle of the distribution includes a smaller baseline distance (and therefore smaller difference in raw scores) than 10 percent of the area (students) near either end of the curve.

In the remainder of the chart, several widely used systems of standard

scores are included. First are the original standard scores, or *z*-scores. The numbers in this scale are the same as those on the baseline of the graph except that the term σ has been omitted. As illustrated above, anyone can translate test scores into *z*-scores by setting the mean for the group equal to 0.0 and the standard deviation equal to 1. The formula for *z* listed on page 21 can be used in translating raw scores into standard-score equivalents. *T*-scores (one of the most widely used systems) equate the mean of the raw-score distribution with 50 and the standard deviation with 10. Thus, a *z*-score of 1.5 is equivalent to a *T*-score of 65 ($1\frac{1}{2}$ standard-deviation units of 10 points above an arbitrary mean of 50). The College Entrance Examination Board avoids the use of decimals by equating the mean raw score on all its tests with 500 points and the standard deviation with 100 points. Hence, the experienced counselor thinks of a College Board score of 550 as one-half standard deviation (of 100 points) above the average (500 points) on the CEEB basic norms. On the Army General Classification Test, a mean of 100 points and a standard deviation of 20 points are used.

Stanine scores (developed by the Air Force and now widely used) derive their name from the fact that they divide the baseline of the normal curve into nine groups; hence "standard nines," or stanines. Except for stanines 1 and 9, at either extreme, these groups are spaced in units of one-half standard deviation. The Wechsler Scale values are used only by psychologists administering this individual intelligence test. The deviation IQ's on the last line, however, are more widely used and should be understood by all teachers.

A large number of interrelationships can be read from this chart. For example, the percentile ranks for any deviation IQ, standard score, *T*-score, and the like can be readily determined. If the baseline is divided into somewhat smaller units and a straight-edge rule is constructed, showing the mean and the values of $+1\sigma$, $+2\sigma$, and the like for a specific distribution of test scores, approximate standard-score equivalents or percentile ranks can be read off for any raw-score value. In interpreting research studies in which stanine scores or *T*-scores are used, the reader can readily translate these measures into the more familiar standard scores or percentile ranks.

APPENDIX D Selected Tables

Table A.1

Equivalent Standard Scores and Percentile Scores in a Normal Distribution

Equivalents for Scores at or Above Mean			Equivalents for Scores at or Below Mean		
DISTANCE OF X FROM MEAN IN SD 's (z -SCORE)	T -SCALED SCORE	PERCENT- TILE SCORE	PERCENT- TILE SCORE	T -SCALED SCORE	DISTANCE OF X FROM MEAN IN SD 's (z -SCORE)
3.0	80	99.9	0.1	20	-3.0
2.9	79	99.8	0.2	21	-2.9
2.8	78	99.7	0.3	22	-2.8
2.7	77	99.6	0.4	23	-2.7
2.6	76	99.5	0.5	24	-2.6
2.5	75	99.4	0.6	25	-2.5
2.4	74	99.2	0.8	26	-2.4
2.3	73	99	1	27	-2.3
2.2	72	99	1	28	-2.2
2.1	71	98	2	29	-2.1
2.0	70	98	2	30	-2.0
1.9	69	97	3	31	-1.9
1.8	68	96	4	32	-1.8
1.7	67	96	4	33	-1.7
1.6	66	95	5	34	-1.6
1.5	65	93	7	35	-1.5
1.4	64	92	8	36	-1.4
1.3	63	90	10	37	-1.3
1.2	62	88	12	38	-1.2
1.1	61	86	14	39	-1.1
1.0	60	84	16	40	-1.0
0.9	59	82	18	41	-0.9
0.8	58	79	21	42	-0.8
0.7	57	76	24	43	-0.7
0.6	56	73	27	44	-0.6
0.5	55	69	31	45	-0.5
0.4	54	66	34	46	-0.4
0.3	53	62	38	47	-0.3
0.2	52	58	42	48	-0.2
0.1	51	54	46	49	-0.1
0.0	50	50	50	50	0.0

Table A.2

Effect of Lengthening a Test on Its Reliability as Predicted
by the Spearman-Brown Formula^a

PRESENT RELIABILITY COEFFICIENT	PREDICTED RELIABILITY WHEN LENGTH OF TEST IS MULTIPLIED BY		
	2	3	4
.70	.823	.875	.903
.72	.837	.885	.911
.74	.851	.895	.919
.76	.864	.905	.927
.78	.876	.915	.934
.80	.888	.923	.941
.82	.901	.932	.948
.84	.913	.940	.955
.86	.925	.948	.960
.88	.936	.956	.967
.90	.947	.964	.973
.92	.958	.972	.979
.94	.969	.979	.984

^a The Spearman-Brown formula is: $r_n = \frac{nr}{1 + (n-1)r}$.

The meanings of symbols are as follows: r = known reliability coefficient; n = number of times the test whose reliability is to be estimated is longer than the one whose reliability is known; r_n = estimated reliability coefficient for test of increased length.

Table A.3
Computation of Mean and Standard Deviation
 (Data: Scores on State History Test for School District)

SCORE INTERVAL	<i>f</i>	<i>d</i>	<i>fd</i>	<i>fd</i> ²	DIRECTIONS FOR COMPUTING MEAN AND SD
96-98	11	+ 8	+ 88	704	1. In each row, multiply the <i>f</i> and <i>d</i> values and enter the products in the <i>fd</i> column.
93-95	21	+ 7	+147	1029	
90-92	30	+ 6	+180	1080	
87-89	25	+ 5	+125	625	
84-86	29	+ 4	+116	464	
81-83	35	+ 3	+105	315	2. In each row, multiply the <i>f</i> and the <i>fd</i> values and enter the products in the <i>fd</i> ² column.
78-80	40	+ 2	+ 80	160	
75-77	40	+ 1	+ 40	40	3. Add the <i>fd</i> column to obtain Σfd . Add the <i>fd</i> ² column to obtain Σfd^2 .
72-74	48	0	(+881)		
69-71	50	- 1	- 50	50	4. Find the correction (in intervals) $c \text{ (correction)} = \frac{\Sigma fd}{N} = \frac{82}{500} = .164$
66-68	40	- 2	- 80	160	
63-65	32	- 3	- 96	288	5. Substitute values obtained in formulas for the mean and standard deviation.
60-62	28	- 4	-112	448	
57-59	26	- 5	-130	650	
54-56	17	- 6	-102	612	
51-53	12	- 7	- 84	588	
48-50	8	- 8	- 64	512	
45-47	3	- 9	- 27	243	
42-44	2	-10	- 20	200	
39-41	2	-11	- 22	242	
36-38	1	-12	- 12	144	
<i>N</i> = 500			(-799) + 82	8554	
			(Σfd) (Σfd^2)		

$$M \text{ (Mean)} = AM + (c)(i) = 73 + (.164)(3) = 73 + .492 = 73.492$$

$$SD = i \sqrt{\frac{\Sigma fd^2}{N} - c^2} = 3 \times \sqrt{\frac{8554}{500} - (.164)^2} = 3\sqrt{17.1080 - .0269} = 3\sqrt{17.0811}$$

$$= 3 (4.133) = 12.399 = 12.4$$

AM is midpoint of the score interval chosen as the arbitrary origin.

		SCORES ON ODD-NUMBERED QUESTIONS (X VARIABLE)																			
		18-20	21-23	24-26	27-29	30-32	33-35	36-38	39-41	42-44	45-47	48-50		f_y	d_y	fd_y	fd_y^2	$\Sigma x'y'$			
																		+	-		
SCORES ON EVEN-NUMBERED QUESTIONS (Y VARIABLE)	48-50	-25	-20	-15	-10	-5		5	10	15	20	25	30	35	10	+5	50	250	220		
	45-47	-20	-16	-12	-8	-4		4	8	12	16	20	24	28	54	+4	216	864	820		
	42-44	-15	-12	-9	-6	-3		3	6	9	12	15	18	21	52	+3	156	468	459		
	39-41	-10	-8	-6	-4	-2		2	4	6	8	10	12	14	78	+2	156	312	284		
	36-38	-5	-4	-3	-2	-1		1	2	3	4	5	6	7	83	+1	83	83	95	-2	
	33-35				1	14	55	13	2						85	0	×	×	×	×	
	30-32	5	4	3	2	1		-1	-2	-3	-4	-5	-6	-7	62	-1	-62	62	46	-1	
	27-29	10	8	6	5	4	3	2	1						45	-2	-90	180	146		
	24-26	15	12	9	6	3	2	1							21	-3	-63	189	159		
	21-23	20	18	12	8	4	3	2	1						6	-4	-24	96	68		
	18-20	25	20	15	10	5	4	3	2	1					4	-5	-20	100	80		
	f_x	2	4	19	41	58	95	93	72	56	48	12			$N=500$		$\Sigma fd_y=402$	2604	$\Sigma x'y'=2374$		
	d_x	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7							
	fd_x	-10	-16	-57	-82	-58	×	93	144	168	192	60					$\Sigma fd_x=434$				
	fd_x^2	50	64	171	164	58	×	93	288	504	768	300					$\Sigma fd_x^2=2460$				
	$\Sigma x'y'$	+					×	101	264	537	716	260					$\Sigma x'y'=2374$				
	$\Sigma x'y'$	-				-2	×	-1													

Fig. A.2 Computation of a Reliability Coefficient (split-halves method) by the Use of the Pearson Product-Moment Method of Correlation.

COMPUTATION OF PEARSON r BY THE FORMULA FOR GROUPED DATA The usual method of computing r from a scatter diagram involves the solution of the following formula:

$$r = \frac{\frac{\Sigma x'y'}{N} - c'_x c'_y}{SD'_x SD'_y},$$

in which N is the number of cases; c'_x is the correction value (in interval units) used in computing the mean of x ; c'_y is the corresponding value for variable y ; SD'_x is the standard deviation for x (in interval-units); SD'_y is the corresponding value for y ; and x' and y' are deviations from AM (assumed mean) in terms of interval units. The reader has already met most of these terms and symbols since they were used in Table 2.2 in the computation of the mean and standard deviation. The new terms (x' and y') are similar to z_x and z_y , except that x' and y' are deviations computed from an assumed

mean (in terms of interval units), rather than deviations from the actual mean (in terms of SD units).

The reader will recognize that the first term in the numerator is similar to the standard-score or z -score formula for r , used in Table 3.3; that is, it is the average of the products of paired deviations. The other terms must be introduced into this more elaborate formula because we are no longer working with standard scores. The second term in the numerator ($c'_x c'_y$) is necessary because we are working with deviations from an assumed mean, rather than deviations from the actual mean. The denominator ($SD'_x SD'_y$) is necessary since the x' values and y' values are not in standard-score form.

Since r involves a measure of relationship and its value does not depend on the size of the original scores, *the corrections and the standard deviations are all left in interval units* to simplify computation. As in the short methods of computing the mean and standard deviation, all scores in a single square (cell) of the scatter-diagram are treated as if they fell at the midpoint of the interval.

Application of the formula above involves seven major steps (illustrated in Figure A.2 and the Computing Guide).

Computation Guide for the Pearson Product-Moment Method of Computing r

1. Tally each pair of scores in a correlation table or scatter diagram (Fig. A.2). Write in each cell the total number of tallies for that cell. Add the frequencies (number of tallies) for each row and enter in the f_y column. Add the frequencies for each column and enter in the f_x row. Total the f_y and f_x values to obtain N .
2. Following the short method of computing the mean (shown in Tables 2.2 and 3.3), compute for both the x and y variables the Σfd value and the correction (in intervals). A subscript is used to denote the Σfd value and the correction for each variable; for example, the Σfd value for x is written Σfd_x . Correction for x is written c'_x .
3. Following the short method for computing the standard deviation (shown in Table A.3), compute the Σfd^2 value and the standard deviation (SD) for each variable, leaving the SD in terms of intervals.
4. Compute the average product moment, $\frac{\Sigma x'y'}{N}$, using the following procedures:
 - a. The frequency in each cell of the scatter diagram is multiplied by the small numeral printed in the cell. This numeral indicates the product moment for that cell (a product of d_x , representing the distance or deviation in intervals from the x axis, and d_y , representing the distance in intervals from the y axis). In this way, the $x'y'$ product for each cell is obtained.

- b. The positive $x'y'$ products and the negative $x'y'$ products are totaled for each row. The partial sums for each row are added algebraically to obtain $\Sigma x'y'$.
 - c. As a cross-check, the positive and negative $x'y'$ products are totaled for each *column*. The partial sums for each column are added algebraically to obtain $\Sigma x'y'$ (which should verify the value obtained above).
 - d. The verified $\Sigma x'y'$ value is divided by N (the number of cases).
5. The product of the corrections (in intervals) for x and y ($c'_x c'_y$) is computed and subtracted from $\frac{\Sigma x'y'}{N}$ to obtain the numerator for the formula.
6. The product of the two standard deviations (in interval units) is computed to obtain the denominator for the formula.
7. The final division of numerator by denominator is made to obtain the value of r . If the sign is negative, the relationship is an inverse one.

APPENDIX E

A Glossary of 100 Measurement Terms

by Roger T. Lennon¹

This glossary of technical terms used in educational and psychological measurement is primarily for persons with limited training in measurement, rather than for the specialist. The terms defined are the more common or basic ones such as occur in test manuals and simple research reports. In the definitions, niceties of usage have sometimes been sacrificed for the sake of brevity and, it is hoped, clarity.

The definitions are based on study of the definitions and usages of the various terms in about a dozen widely used textbooks in educational and psychological measurement and statistics, and in both general and specialized dictionaries. There is not complete uniformity among writers in the measurement field with respect to the usage of certain technical terms; in cases of varying usage, either these variations are noted or the definition offered is the one that the writer judges to represent the "best" usage.

academic aptitude. The combination of native and acquired abilities that is needed for school work; likelihood of success in mastering academic work, as estimated from measures of the necessary abilities. (Also called *scholastic aptitude*.)

accomplishment quotient (AQ). The ratio of educational age to mental age; $EA \div MA$. (Also called *achievement quotient*.)

achievement age. The age for which a given achievement test score is the real or estimated average. (Also called *educational age* or *subject age*.) If the achievement age corresponding to a score of 36 on a reading test is 10 years, 7 months (10-7), this means that pupils 10 years, 7 months achieve, on the average, a score of 36 on that test.

achievement test. A test that measures the extent to which a person has "achieved" something—acquired certain information or mastered certain skills, usually as a result of specific instruction.

¹ Published as *Test Service Notebook No. 13* (New York: Harcourt, Brace & World, Inc.). Compiled with the assistance of Claude F. Bridges, John C. Marriott, Frances E. Crook, and Blythe C. Mitchell, Division of Test Research and Service. Reprinted with the permission of Harcourt, Brace & World, Inc.

- age equivalent.** The age for which a given score is the real or estimated average score.
- age norms.** Values representing typical or average performance for persons of various age groups.
- age-grade table.** A table showing the number or per cent of pupils of various ages in each grade; a distribution of the ages of pupils in successive grades.
- alternate-form reliability.** The closeness of correspondence, or correlation, between results on alternate (*i.e.* equivalent or parallel) forms of a test; thus, a measure of the extent to which the two forms are consistent or reliable in measuring whatever they do measure, assuming that the examinees themselves do not change in the abilities measured between the two testings. (See RELIABILITY, RELIABILITY COEFFICIENT, STANDARD ERROR.)
- aptitude.** A combination of abilities and other characteristics, whether native or acquired, known or believed to be indicative of an individual's ability to learn in some particular area. Thus, "musical aptitude" would refer broadly to that combination of physical and mental characteristics, motivational factors, and conceivably other characteristics, which is conducive to acquiring proficiency in the musical field. Some exclude motivational factors, including interests, from the concept of "aptitude," but the more comprehensive use seems preferable. The layman may think of "aptitude" as referring only to some inborn capacity; the term is no longer so restricted in its psychological or measurement usage.
- arithmetic mean.** The sum of a set of scores divided by the number of scores. (Commonly called *average*, *mean*.)
- average.** A general term applied to measures of central tendency. The three most widely used averages are the *arithmetic mean*, the *median*, and the *mode*.
- battery.** A group of several tests standardized on the same population, so that results on the several tests are comparable. Sometimes loosely applied to any group of tests administered together, even though not standardized on the same subjects.
- ceiling.** The upper limit of ability measured by a test.
- class analysis chart.** A chart, usually prepared in connection with a battery of achievement tests, that shows the relative performance of members of a class on the several parts of the battery.
- coefficient of correlation (r).** A measure of the degree of relationship, or "going-togetherness," between two sets of measures for the same group of individuals. The correlation coefficient most frequently used in test development and educational research is that known as the *Pearson (Pearsonian) r* , so named for Karl Pearson, originator of the method, or as the *product-moment r* , to denote the mathematical basis of its calculation. Unless otherwise specified, "correlation" usually means the product-moment correlation coefficient, which ranges from .00, denoting complete absence of relationship, to 1.00, denoting perfect correspondence, and may be either positive or negative.

- completion item.** A test question calling for the completion (filling in) of a phrase, sentence, etc., from which one or more parts have been omitted.
- correction for guessing.** A reduction in score for wrong answers, sometimes applied in scoring true-false or multiple-choice questions. Many question the validity or usefulness of this device, which is intended to discourage guessing and to yield more accurate rankings of examinees in terms of their true knowledge. Scores to which such corrections have been applied—e.g., rights minus wrongs, or rights minus some fraction of wrongs—are often spoken of as “corrected for guessing” or “corrected for chance.”
- correlation.** Relationship or “going-togetherness” between two scores or measures; tendency of one score to vary concomitantly with the other, as the tendency of students of high IQ to be above average in reading ability. The existence of a strong relationship—i.e., a high correlation—between two variables does not necessarily indicate that one has any causal influence on the other. (See COEFFICIENT OF CORRELATION.)
- criterion.** A standard by which a test may be judged or evaluated; a set of scores, ratings, etc., that a test is designed to predict or to correlate with. (See VALIDITY.)
- decile.** Any one of the nine percentile points (scores) in a distribution that divide the distribution into ten equal parts; every tenth percentile. The first decile is the 10th percentile, the ninth decile the 90th percentile, etc.
- deviation.** The amount by which a score differs from some reference value, such as the mean, the norm, or the score on some other test.
- deviation IQ.** See INTELLIGENCE QUOTIENT.
- diagnostic test.** A test used to “diagnose,” that is, to locate specific areas of weakness or strength, and to determine the nature of weaknesses or deficiencies; it yields measures of the components or sub-parts of some larger body of information or skill. Diagnostic achievement tests are most commonly prepared for the skill subjects—reading, arithmetic, spelling.
- difficulty value.** The per cent of some specified group, such as students of a given age or grade, who answer an item correctly.
- discriminating power.** The ability of a test item to differentiate between persons possessing much of some trait and those possessing little.
- distractor.** Any of the incorrect choices in a multiple-choice or matching item.
- distribution (frequency distribution).** A tabulation of scores from high to low, or low to high, showing the number of individuals that obtain each score or fall in each score interval.
- educational age (EA).** See ACHIEVEMENT AGE.
- equivalent form.** Any of two or more forms of a test that are closely parallel with respect to the nature of the content and the difficulty of the items included, and that will yield very similar average scores and measures of variability for a given group.
- error of measurement.** See STANDARD ERROR.

extrapolation. In general, any process of estimating values of a function beyond the range of available data. As applied to test norms, the process of extending a norm line beyond the limits of actually obtained data, in order to permit interpretation of extreme scores. This extension may be done mathematically by fitting a curve to the obtained data or, as is more common, by less rigorous methods, usually graphic. See Fig. 1. Considerable judgment on the test maker's part enters into any extrapolation process, which means that extrapolated norm values are likely to be to some extent arbitrary.

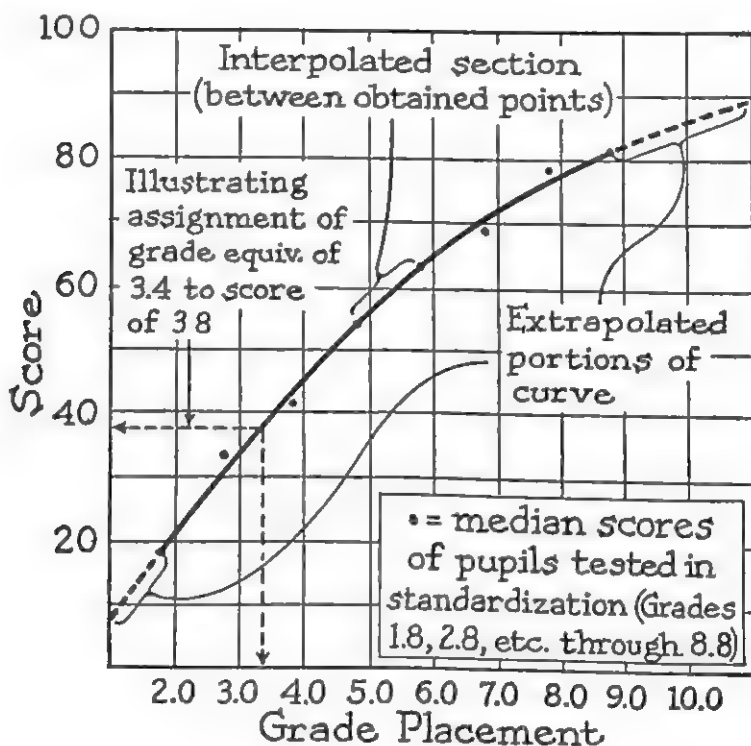


Fig. 1

factor. In mental measurement, a hypothetical trait, ability or component of ability, that underlies and influences performance on two or more tests, and hence causes scores on the tests to be correlated. The term "factor" strictly refers to a theoretical variable, derived by a process of *factor analysis*, from a table of intercorrelations among tests; but it is also commonly used to denote the psychological interpretation given to the variable—i.e., the mental trait assumed to be represented by the variable, as verbal ability, numerical ability, etc.

factor analysis. Any of several methods of analyzing the intercorrelations among a set of variables such as test scores. Factor analysis attempts to account for the interrelationships in terms of some underlying "factors," preferably fewer in number than the original variables; and it reveals how

forced-choice item. Broadly, any multiple-choice item in which the examinee is *required* to select one or more of the given choices. The term is best used to denote a special type of multiple-choice item, in which the options, or choices, are (1) of equal "preference value"—i.e., chosen equally often by a typical group, but (2) of differential discriminating ability—i.e., such that one of the options discriminates between persons high and low on the factor that this option measures, while the other options do not.

grade equivalent. The grade level for which a given score is the real or estimated average.

group test. A test that may be administered to a number of individuals at the same time by one examiner.

intelligence quotient (IQ). Originally, the ratio of a person's mental age to his chronological age $\left(\frac{MA}{CA} \right)$ or, more precisely, especially for older

The following table shows the classification of IQ's offered by Terman and Merrill for the Stanford-Binet test, indicating the per cent of persons in a normal population who fall in each classification. This table is roughly applicable to tests yielding IQ's having standard deviations of about 16 points (not all do). It is important to bear in mind that any such table is arbitrary, for there are no inflexible lines of demarcation between "feeble-minded" and "borderline," etc.

<u>Classification</u>	<u>IQ</u>	<u>Per cents of all persons</u>
Near genius or genius	140 and above	1
Very superior	130-139	2.5
Superior	120-129	8
Above average	110-119	16
Normal or average	90-109	45
Below average	80-89	16
Dull or borderline	70-79	8
Feeble-minded: moron,	60-69	2.5
imbecile, idiot	59 and below	1

interpolation. In general, any process of estimating intermediate values between two known points. As applied to test norms, it refers to the procedure used in assigning interpreted values (e.g., grade or age equivalents) to scores between the successive average scores actually obtained in the standardization process. In reading norm tables, it is necessary at times to *interpolate* to obtain a norm value for a score between scores given in the table; e.g., in the table given here, an age value of 12-5 would be assigned, by interpolation, to a score of 118. See Fig. 1. under EXTRA-

Score	Age Equiv.
120	12-6
115	12-4
110	12-2

inventory test. As applied to achievement tests, a test that attempts to cover rather thoroughly some relatively small unit of specific instruction or training. The purpose of an inventory test, as the name suggests, is more in the nature of a "stock-taking" of an individual's knowledge or skill than an effort to measure in the usual sense. The term sometimes denotes a type of test used to measure achievement status prior to instruction. Many personality and interest questionnaires are designated "inventories," since they appraise an individual's status in several personal characteristics, or his level of interest in a variety of types of activities.

item. A single question or exercise in a test.

item analysis. The process of evaluating single test items by any of several methods. It usually involves determining the difficulty value and the discriminating power of the item, and often its correlation with some criterion.

Kuder-Richardson formula(s). Formulas for estimating the reliability of a test from information about the individual items in the test, or from the mean score, standard deviation, and number of items in the test. Because the Kuder-Richardson formulas permit estimation of reliability from a single administration of a test, without the labor involved in dividing the test into halves, their use has become common in test development. The Kuder-Richardson formulas are not appropriate for estimating the reliability of speeded tests.

machine-scorable (machine-scored) test. A test that may be scored by means of a machine. Ordinarily, the term refers to a test adapted for scoring on the International Test Scoring Machine, manufactured by International Business Machines Corporation. In taking tests that are to be scored on this machine, the examinee records his answers on separate answer sheets with a special electrographic pencil. These pencil marks are electrically conductive, and current flowing through them may be read on a suitably calibrated dial as a test score. The machine distinguishes, by means of appropriate keys, between right and wrong answers, and can combine groups of responses in order to yield total or part scores, weighted scores, or corrected scores.

matching item. A test item calling for the correct association of each entry in one list with an entry in a second list.

mean. See ARITHMETIC MEAN.

median. The middle score in a distribution; the 50th percentile; the point that divides the group into two equal parts. Half of the group of scores fall below the median and half above it.

mental age (MA). The age for which a given score on an intelligence test is average or normal. If a score of 55 on an intelligence test corresponds to a mental age of 6 years, 10 months, then 55 is presumably the average score that would be made by an unselected group of children 6 years, 10 months of age.

modal age. That age or age range which is most typical or characteristic of pupils of specified grade placement.

modal-age norms. Norms based on the performance of pupils of modal age for their respective grades, which are thus free of the distorting influence of under-age or over-age pupils.

mode. The score or value that occurs most frequently in a distribution.

multiple-choice item. A test item in which the examinee's task is to choose the correct or best answer from several given answers, or options.

multiple-response item. A special type of multiple-choice item in which two or more of the given choices may be correct.

N. The symbol commonly used to represent the number of cases in a distribution, study, etc.

normal distribution. A distribution of scores or measures that in graphic form has a distinctive bell-shaped appearance. Figure 2 shows such a graph of a normal distribution, known as a *normal curve* or *normal probability curve*. In a normal distribution, scores or measures are distributed symmetrically about the mean, with as many cases at various distances above the mean as at equal distances below it, and with cases concentrated near the average and decreasing in frequency the further one departs from the average, according to a precise mathematical equation. The assumption that mental and psychological characteristics are distributed normally has been very useful in much test development work.

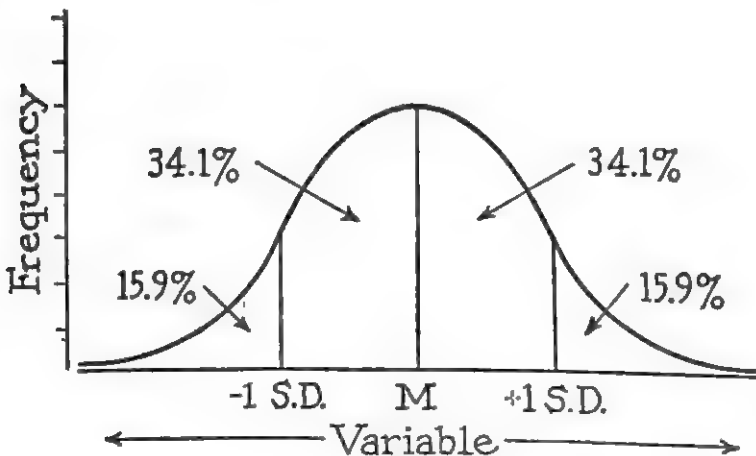


Fig. 2

- norm line.** A smooth curve drawn through the mean or median scores of successive age or grade groups, or through percentile points for a single group. See Fig. 1 under **EXTRAPOLATION**.
- norms.** Statistics that describe the test performance of specified groups, such as pupils of various ages or grades in the standardization group for a test. Norms are often assumed to be representative of some larger population, as of pupils in the county as a whole. Norms are descriptive of average, typical, or mediocre performance; they are not to be regarded as standards, or as desirable levels of attainment. Grade, age, and percentile are the most common types of norms.
- objective test.** A test in the scoring of which there is no possibility of difference of opinion among scorers as to whether responses are to be scored right or wrong. It is contrasted with a "subjective" test—e.g., the usual essay examination to which different scorers may assign different scores, ratings, or grades.
- omnibus test.** A test (1) in which items measuring a variety of mental operations are all combined into a single sequence rather than being grouped together by type of operation, and (2) from which only a single score is derived, rather than separate scores for each operation or function. Omnibus tests make for simplicity of administration: one set of directions and one over-all time limit usually suffice. *Otis Quick-Scoring Mental Ability Tests: Beta or Gamma Tests* are omnibus-type tests, as distinguished from tests such as *Terman-McNemar Test of Mental Ability* or *Pintner General Ability Tests: Verbal*, in which the items measuring various operations are grouped together, each with its own set of directions.
- percentile (P).** A point (score) in a distribution below which falls the per cent of cases indicated by the given percentile. Thus the 15th percentile denotes the score or point below which 15 per cent of the scores fall. "Percentile" has nothing to do with the per cent of correct answers an examinee has on a test.
- percentile rank.** The per cent of scores in a distribution equal to or lower than the score corresponding to the given rank.
- performance test.** As contrasted with *paper-and-pencil test*, a test requiring motor or manual response on the examinee's part, generally but not always involving manipulation of concrete equipment or materials. *Cornell-Coxe Performance Ability Scale*, *Arthur Point Scale of Performance Tests*, and *Bennett Hand-Tool Dexterity Test* are performance tests, in this sense. "Performance test" is also used in another sense, to denote a test that is actually a work-sample, and in this sense it may include paper-and-pencil tests, as, for example, a test in accountancy, or in taking shorthand, or in proofreading, where no materials other than paper and pencil may be required, but where the test response is identical with the behavior about which information is desired.
- personality test.** A test intended to measure one or more of the non-intellective aspects of an individual's mental or psychological make-up. Personality tests include the so-called *personality inventories* or *adjustment inventories* (e.g., *Heston Personal Adjustment Inventory*, *Bernreuter Personality Inventory*, *Bell Adjustment Inventory*) which seek to measure a person's

status on such traits as dominance, sociability, introversion, etc., by means of self-descriptive responses to a series of questions; *rating scales* (e.g., *Haggerty-Olson-Wickman Behavior Rating Schedules*) which call for rating, by one's self or another, of the extent to which a subject possesses certain characteristics; *situation tests* in which the individual's behavior in simulated life-like situations is observed by one or more judges, and evaluated with reference to various personality traits; and *opinion or attitude inventories* (e.g., *Allport-Vernon Study of Values*). Some writers also classify interest inventories as personality tests.

power test. A test intended to measure level of performance rather than speed of response; hence one in which there is either no time limit or a very generous one.

practice effect. The influence of previous experience with a test on a later administration of the same test or a similar test; usually, an increase in the score on the second testing, attributed to increased familiarity with the directions, kinds of questions, etc. Practice effect is greatest when the interval between testings is small, when the materials in the two tests are very similar, and when the initial test-taking represents a relatively novel experience for the subjects.

probable error. See STANDARD ERROR.

product-moment coefficient. See COEFFICIENT OF CORRELATION.

profile. A graphic representation of the results on several tests, for either an individual or a group, when the results have been expressed in some uniform or comparable terms. This method of presentation permits easy identification of areas of strength or weakness.

projective technique (projective method). A method of personality study in which the subject responds as he chooses to a series of stimuli such as ink-blot, pictures, unfinished sentences, etc. So called because of the assumption that under this free-response condition the subject "projects" into his responses manifestations of personality characteristics and organization that can, by suitable methods, be scored and interpreted to yield a description of his basic personality structure. The *Rorschach* (ink-blot) *Technique* and the *Murray Thematic Apperception Test* are the most commonly used projective methods.

prognosis (prognostic) test. A test used to predict future success or failure in a specific subject or field.

quartile. One of three points that divide the cases in a distribution into four equal groups. The lower quartile, or 25th percentile, sets off the lowest fourth of the group; the middle quartile is the same as the 50th percentile, or median; and the third quartile, or 75th percentile, marks off the highest fourth.

r. See COEFFICIENT OF CORRELATION.

random sample. A sample of the members of a population drawn in such a way that every member of the population has an equal chance of being included—that is, drawn in a way that precludes the operation of bias or selection. The purpose in using a sample thus free of bias is, of course,

that the sample be fairly "representative" of the total population, so that sample findings may be generalized to the population. A great advantage of random samples is that formulas are available for estimating the expected variation of the sample statistics from their true values in the total population; in other words, we know how precise an estimate of the population value is given by a random sample of any given size.

range. The difference between the lowest and highest scores obtained on a test by some group.

raw score. The first quantitative result obtained in scoring a test. Usually the number of right answers, number right minus some fraction of number wrong, time required for performance, number of errors, or similar direct, unconverted, uninterpreted measure.

readiness test. A test that measures the extent to which an individual has achieved a degree of maturity or acquired certain skills or information needed for undertaking successfully some new learning activity. Thus a *reading readiness test* indicates the extent to which a child has reached a developmental stage where he may profitably begin a formal instructional program in reading.

recall item. An item that requires the examinee to supply the correct answer from his own memory or recollection, as contrasted with a *recognition item*, in which he need only identify the correct answer.

e.g., "Columbus discovered America in the year ? "

is a recall item, whereas

"Columbus discovered America in *a* 1425 *b* 1492 *c* 1520 *d* 1546"

is a recognition item.

recognition item. An item requiring the examinee to recognize or select the correct answer from among two or more given answers. See **RECALL ITEM**.

reliability. The extent to which a test is consistent in measuring whatever it does measure; dependability, stability, relative freedom from errors of measurement. Reliability is usually estimated by some form of *reliability coefficient* or by the *standard error of measurement*.

reliability coefficient. The coefficient of correlation between two forms of a test, between scores on repeated administrations of the same test, or between halves of a test, properly corrected. These three coefficients measure somewhat different aspects of reliability but all are properly spoken of as reliability coefficients. See **ALTERNATE-FORM RELIABILITY**, **SPLIT-HALF COEFFICIENT**, **TEST-RETEST COEFFICIENT**, **KUDER-RICHARDSON FORMULA(S)**.

representative sample. A sample that corresponds to or matches the population of which it is a sample with respect to characteristics important for the purposes under investigation—e.g., in an achievement test norm sample, proportion of pupils from each state, from various regions, from segregated and non-segregated schools, etc.

scholastic aptitude. See **ACADEMIC APTITUDE**.

skewness. The tendency of a distribution to depart from symmetry or balance around the mean.

sociometry. Measurement of the interpersonal relationships prevailing among the members of a group. By means of sociometric devices, e.g., the *sociogram*, an attempt is made to discover the patterns of choice and rejection among the individuals making up the group—which ones are chosen most often as friends or leaders (“stars”), which are rejected by others (“isolates”), how the group subdivides into clusters or cliques, etc.

Spearman-Brown formula. A formula giving the relationship between the reliability of a test and its length. The formula permits estimation of the reliability of a test lengthened or shortened by any amount, from the known reliability of a test of specified length. Its most common application is in the estimation of reliability of an entire test from the correlation between two halves of the test (*split-half reliability*).

split-half coefficient. A coefficient of reliability obtained by correlating scores on one half of a test with scores on the other half. Generally, but not necessarily, the two halves consist of the odd-numbered and the even-numbered items.

standard deviation (S.D.). A measure of the variability or dispersion of a set of scores. The more the scores cluster around the mean, the smaller the standard deviation.

standard error (S.E.). An estimate of the magnitude of the “error of measurement” in a score—that is, the amount by which an obtained score differs from a hypothetical true score. The standard error is an amount such that in about two-thirds of the cases the obtained score would not differ by more than one standard error from the true score. The *probable error* (P.E.) of a score is a similar measure, except that in about half the cases the obtained score differs from the true score by not more than one probable error. The probable error is equal to about two-thirds of the standard error. The larger the probable or the standard error of a score, the less reliable the measure.

standard score. A general term referring to any of a variety of “transformed” scores, in terms of which raw scores may be expressed for reasons of convenience, comparability, ease of interpretation, etc.

The simplest type of standard score is that which expresses the deviation of an individual's raw score from the average score of his group in relation to the standard deviation of the scores of the groups. Thus:

$$\text{Standard score (z)} = \frac{\text{raw score (X)} - \text{mean (M)}}{\text{standard deviation (S.D.)}}$$

By multiplying this ratio by a suitable constant and by adding or subtracting another constant, standard scores having any desired mean and standard deviation may be obtained. Such standard scores do not affect the relative standing of the individuals in the group nor change the shape of the original distribution.

More complicated types of standard scores may yield distributions differing in shape from the original distribution; in fact, they are sometimes used for precisely this purpose. *Normalized standard scores* and *K-scores* (as used in *Stanford Achievement Test*) are examples of this latter group.

- standardized test (standard test).** A systematic sample of performance obtained under prescribed conditions, scored according to definite rules, and capable of evaluation by reference to normative information. Some writers restrict the term to tests having the above properties, whose items have been experimentally evaluated, and/or for which evidences of validity and reliability are provided.
- stanine.** One of the steps in a nine-point scale of normalized standard scores. The stanine (short for *standard-nine*) scale has values from 1 to 9, with a mean of 5, and a standard deviation of 2.
- stencil key.** A scoring key which, when positioned over an examinee's responses either in a test booklet or, more commonly, on an answer sheet, permits rapid identification and counting of all right answers. Stencil keys may be perforated in positions corresponding to positions of right answers, so that only right answers show through when the keys are in place; or they may be transparent, with positions of right answers identified by circles, boxes, etc., printed on the key.
- strip key.** A scoring key arranged so that the answers for items on any page or in any column of the test appear in a strip or column that may be placed alongside the examinee's responses for easy scoring.
- survey test.** A test that measures general achievement in a given subject or area, usually with the connotation that the test is intended to measure group status, rather than to yield precise measures of individuals.
- test-retest coefficient.** A type of reliability coefficient obtained by administering the same test a second time after a short interval and correlating the two sets of scores.
- true-false item.** A test question or exercise in which the examinee's task is to indicate whether a given statement is true or false.
- true score.** A score entirely free of errors of measurement. True scores are hypothetical values never obtained by testing, which always involves some measurement error. A true score is sometimes defined as the average score of an infinite series of measurements with the same or exactly equivalent tests, assuming no practice effect or change in the examinee during the testings.
- validity.** The extent to which a test does the job for which it is used. Validity, thus defined, has different connotations for various kinds of tests and, accordingly, different kinds of validity evidence are appropriate for them. For example:
- (1) The validity of an achievement test is the extent to which the content of the test represents a balanced and adequate sampling of the outcomes (knowledge, skills, etc.) of the course or instructional program it is intended to cover (*content, face, or curricular validity*). It is best evidenced by a comparison of the test content with courses of study, instructional materials and statements of instructional goals, and by critical analysis of the processes required in responding to the items.
 - (2) The validity of an aptitude, prognostic, or readiness test is the extent to which it accurately indicates future learning success in the area

for which it is used as a predictor (*predictive validity*). It is evidenced by correlations between test scores and measures of later success.

(3) The validity of a personality test is the extent to which the test yields an accurate description of an individual's personality traits or personality organization (*status validity*). It may be evidenced by agreement between test results and other types of evaluation, such as ratings or clinical classification, but only to the extent that such criteria are themselves valid.

The traditional definition of validity as "the extent to which a test measures what it is supposed to measure," seems less satisfactory than the above, since it fails to emphasize that the validity of a test is always specific to the purposes for which the test is used, and that different kinds of evidence are appropriate for appraising the validity of various types of tests.

Validity of a test *item* refers to the discriminating power of the item—its ability to distinguish between persons having much and those having little of some characteristic.

Index

- Ability
developed versus innate, 181
tests of, 154
see also Aptitude
- Achievement
definition of, 154, 182
interindividual differences in, 459
intraindividual differences in, 459
- Achievement tests, 437 ff., 496–497
development of, 428–432
for elementary and junior high school, 440–444, 496
for high school, 444–450, 496–497
performance, 402
results from, 450–454
selection of, 437 ff.
uses of, 432–437
- Adams, Georgia Sachs, 419
- Adjustment. *See* Personal-social adjustment
- Adkins, D. C., 230
- Administration
of evaluation program, 488, 499–504
problems in school, 9–14
of tests, 166, 500–503
- Aesthetic Perception Test*, 209
- Age
chronological, 62
educational, 56
mental, 56, 184
reading, 56
- Age scores, 55–56
- Allport, G. W., 241, 297
- Alschuler, Rose H., 316
- American Psychological Association, 5, 167, 184, 237
- American Textbook Publishers Institute, 61
- Analysis, as educational goal, 365, 386–388
- Anastasi, Anne, 40, 312
- "Anchor test," 421, 530, 531
- Anderson, H. A., 423
- Anderson, H. C., 243
- Anderson Chemistry Test*, 369
- Anecdotal record, 273, 508
- Annual Review of Psychology*, 175
- Anticipated Achievement Grade Placements (AAGP), 440, 450, 465, 466, 468, 469
- Application, as educational goal, 365, 383–386
- Aptitude
as combination of abilities, 182–183
definition of, 181
interests and, 230–232
learning difficulties and, 476
measurement of, 181 ff.
unitary, 182–183
- Aptitude tests, 58, 115, 122, 130, 154, 495–496
in art, 209
batteries, 183, 194–204
compared with multiscore intelligence tests, 198–203
interpretation of, 219–220
for clerical skills, 208
development of, 183–186
for foreign languages, 211
interpretation of results from, 214–222
for manual dexterity, 205, 208
for mathematics, 211

- Aptitude tests (*Cont.*)
 for music, 209–210
 paper and pencil, 205
 performance, 205
 power vs. speed, 200
 predictions resulting from, 222
 prognostic, 210–212, 234
 purpose of, 212–214
 results from, versus achievement test
 results, 450–452
 for shorthand, 211–212
 special, 204–210
 use of results from, 220–221
- Area transformation, 38
- Arithmetic
 diagnostic tests in, 472, 478
 grade norms in, 51
- Army Alpha*, 184, 185
- Army Beta*, 185
- Army, Clara Brown, 176
- Art, aptitude in, 209
- Arthur Point Scale of Performance Tests*, 184
- Athletics, evaluations in, 417
- Attitudes
 definition of, 248
 expressed, 250–251
 inventory of, 250
 manifest, 248–249
 role of teacher in observing, 249
 measurement of, 248–253
 scales for determining, 251–253
- Authoritarianism, measurement of, 137
- Autobiography
 interpretation of, 269
 personal-social adjustment and, 268–270
- Average. *See* Mean
- Batteries, test
 achievement, 440–450
 aptitude, 183 ff.
- Behavior
 abnormal, 260–261
 adaptive, 260
 criterion
 direct versus indirect measurement of, 150–151
 intermediate, 105–106
 tests as measures of, 103–106
 tests as predictors of, 105–106
 ultimate, 105–106
 evaluation of, 270–276
 influenced by attitudes, 248–250
 normal, 259
 observation of, 263–265, 270–276
 relevant to personal-social adjustment, 263–264
 traits, 277–278
- Behavior Preference Record*, 493
- Bennett, George K., 222
- Bennett Test of Mechanical Comprehension*, 204
- Billett-Starr Youth Problems Inventory*, 311
- Binet, Alfred, 184
- Blair, Glenn M., 175
- Bloom, Benjamin S., 363, 394
- Bond, Guy L., 175
- Bonney, M. E., 286
- Brownell, William A., 324, 325
- Brueckner, Leo J., 472
- Buhler, Charlotte, 258
- Buros, Oscar K., 174
- Buros Yearbooks, 174, 204, 370, 437, 440, 444
- California Achievement Test (CAT)*, 51, 52, 54, 62, 169, 437, 438, 439, 440, 444, 450, 465, 473
- California F Scale*, 137
- California Personality Test (CPT)*, 302, 308
- California Psychological Inventory (CPI)*, 303, 309
- California Study Methods Survey*, 492
- California Test Bureau, 168
- California Test of Mental Maturity*, 192, 438
- California Tests in Social and Related Sciences*, 345, 433, 492, 493
- Campbell, Donald T., 137
- Carroll, John B., 211
- Cattell, Raymond B., 297, 298, 305, 316

- Checklist
 - definition of, 407
 - problems, 310-311, 498
 - self-rating, 492
 - teachers', 406, 422, 492
 - for test administrators, 503
- Chicago Test of Primary Mental Abilities*, 199, 230
- Clark, Kenneth E., 244
- Clarke, H. Harrison, 176, 419
- Clerical aptitude, 208-209
- Clinical approach, 262, 311-316
 - versus psychometric approach, 261-262, 295-296
- Coefficient
 - of correlation, 74-83
 - of equivalence, 86, 96
 - of reliability, 83-97
 - of validity, 108, 120, 122-123, 145
- College Entrance Examination Board (CEEB), 172, 173, 174, 413, 445
 - tests of, 45, 352, 454
- Committee on Gifted Students, 12
- Completion questions, 334-336
- Comprehension, as educational goal, 365, 378
 - extrapolation and, 381-382
 - interpretation and, 381
 - translation and, 378-380
- Conference
 - parent-teacher, 516-520
 - pupil-teacher, 265-268
- Constructs, *see* Traits
- Converted scores, 17, 28-33
 - compared with perfect, 18
 - methods for obtaining, 20 ff.
 - scaled, 56
 - types of, 44-49
 - use of, 63-65
 - see also* Age scores, Grade scores, Percentile scores, Stanine scores
- Cooperative English Tests*, 372, 446, 525
- Cooperative Foreign Language Tests*, 447
- Cooperative French Tests*, 372
- Cooperative General Achievement Tests*, 446
- Cooperative Mathematics Tests*, 446
- Cooperative Social Studies Tests*, 447
- Cooperative Study in General Education, 432
- Cooperative Test Division, 57, 446
- Corrective instruction, 480-482
 - materials for, 481
 - program for, 482
- Correlation
 - coefficient of, 74-83, 120
 - bi-serial, 354-355
 - Pearson product-moment method of, 77-78, 232, 624-626
 - reliability and, 73-74
 - Spearman rank-difference method of, 75-76
 - interpretation of, 79-83
 - multiple, 547-548
 - prediction and, 179-182, 547-549
 - standard error of estimate and, 82
 - variance and, 82, 132
- Correlation matrix, 132, 133
- Counseling. *See* Guidance
- Crary American History Test*, 433
- Criteria
 - behavior, 103-106, 150-151
 - of evaluation objectives, 391-392
 - for test selection, 157 ff., 324, 376
- Crites, John O., 175, 200, 208, 232, 241, 254, 539
- Cronbach, Lee J., 263, 302, 308
- Cultural background
 - and achievement test, 450
 - and intelligence test, 217-218
- Cumulative-record system
 - characteristics of, 509-510
 - definition of, 507
 - preparation of, 510-511
 - purposes of, 508
 - use by teachers of, 510, 511-512
- D-statistic, 550-552
- Darley, J. G., 232, 236
- Data, measurement and evaluation
 - explanation of, 4
 - from achievement test, 450-454

Data, measurement and evaluation

(Cont.)

- combining and weighting of, 522–523
 - importance of using local, 123
 - interpretation of, 17 ff., 538–543
 - use of expectancy tables in, 121, 122, 546
 - see also Converted scores, Reliability, Validity
 - on personal-social adjustment, 263 ff.
 - relevancy of, 3, 4, 6
 - reporting of, 512–521
 - effectiveness of, 519–521
 - functions of, 512–514
 - types of, 514–519
 - sociometric. *See* Sociometric data
 - student reaction to, 540
 - summarizing and recording of, 507–512
 - use of, in guidance, 534, 541 ff.
- Davis, Allison, 217
- Davis Reading Test*, 162
- Davis-Eells intelligence tests, 217–218
- Decile point, 33
- Decile rank, 33
- Delinquents, 261
- Design Judgment Test*, 209
- Diagnosis. *See* Educational diagnosis
- Diagnostic Analysis, 471
- Diagnostic Examination in Reading Abilities*, 473
- Dictionary of Occupational Titles*, 240
- Dictionary of Psychology*, 181
- Diederich, Paul P., 41, 355, 406, 422
- Differences
 - interindividual, 459
 - intraindividual, 459, 546
- Differential Aptitude Tests*, (DAT), 163, 197, 199, 200, 220, 221, 230, 231, 550, 552
 - compared with *General Aptitude Test Battery*, 200–203
- Differential predictions, 546–549
 - vocational guidance and, 549
- Distribution curve. *See* Normal distribution curve

Doppelt, Jerome E., 222

Drake Musical Aptitude Test, 210*Draw-a-Person-Test*, 314

Dressel, Paul L., 63, 325, 329

Driscoll Play Kit, 314

Dunning, Gordon M., 346

Durost, Walter N., 450

Durrell Analysis of Reading Difficulty, 475

Dvorak, Beatrice, 200

Dyer, Henry S., 7

Ebel, Robert L., 359, 360, 517

Education, objectives of, 363 ff.

Educational diagnosis, 458 ff.

- by group analysis, 478–480

- levels of, 458, 462

- steps in, 463–477

Educational and Psychological Measurement, 175

Educational Testing Service, 99, 135, 168, 176, 360, 413, 446

Edwards Personal Preference Schedule (EPPS), 309, 310

Eells, Kenneth, 217

Ellis, Albert, 301

Elsbree, Willard S., 514

Emotions, learning ability and, 477

Environment

- cultural. *See* Cultural background
- socioeconomic, 556

Equivalence

- coefficient of, 86

- measurement of, 90

Equivalent-forms method, 86

Error

- amount of, 4

- compensating, 68, 71, 72

- in measurement, 15, 65, 69 ff.

- sources of, 4, 69–70

- standard. *See* Standard error

- systematic, 69, 71

Essay writing, evaluation of, 413, 415–416, 423–424

Essential High School Content Battery, 445, 446

Evaluation

- definition of, 5

Evaluation (*Cont.*)

as educational objective, 366, 391–394

effective teaching and, 461

of personal-social adjustment, 257 ff.

problems in, 9–14

program of, 487 ff.

administration of, 499–504

characteristics of, 489–490

disadvantages of, 454–455

functions of, 488–489

guidance and, 488–489

instruction and, 488

objectives of, 491

planning of, 490–491, 495–499

scheduling of, 498–499

supervision of, 488

teacher and, 487

techniques used in, 492–493

relation to instruction, 325–326

relation to measurement, 6

of skills, 401 ff.

teacher's role in, 8–9, 359–360

Evaluation and Adjustment Series, 57, 169, 447, 448

Examinations. *See* Tests

Expectancy chart, 469, 546

F-score, 303

Factor analysis, 132–137, 147, 195, 239

Fifth Mental Measurements Yearbook, 312

Findley, Warren G., 452

Flanagan Aptitude Classification Tests, 221

Frederiksen, Norman, 245

French, John W., 424

Frequency distribution, 22, 29, 30, 42

Frequency polygon, 35, 37

Froehlich, Clifford P., 175

Fruchter, Benjamin, 136

Gage, N. L., 333

Gardner, Eric F., 290

Gates Diagnostic Reading Tests, 474

Gates Reading Readiness Tests, 211

General Aptitude Test Battery (GATB), 200–204, 205, 221, 553
compared with *Differential Aptitude Tests*, 201–203

General Clerical Test, 58, 208

Gerberich, Raymond, 368

Gifted students, 12, 138, 435

Gilmore Oral Reading Test, 475

Goals. *See* Objectives

Goldman, Leo, 175

Gough, Harrison C., 309

Grade placement norms, 20, 165, 439

Grade scores, 43, 50–55

problems in interpreting, 53–55

Grading

centralized control of, 524

comparability of, 524, 528

in honors classes, 525, 527, 528

reliability of, 521–522

standard policy for, 530–531

validity of, 521–522

variations in practices of, 10–11

Graves, Maitland, 209

Greene, Edward B., 248

Gronlund, Norman E., 286, 287

Guidance

educational, 13, 119, 197, 543, 545
evaluation program and, 488–489, 494, 495

functions of, 534

role of teacher in, 535, 536–538

role of trained personnel in, 536

use of interest inventories in, 228, 236, 239

use of test data in, 541–543

vocational, 13, 221, 543, 545

personality inventory in, 301

study of personality in, 261–262

use of achievement tests in, 437

use of differential predictions in, 546–547

use of interest inventory in, 246–248

use of profile in, 550

Guilford, J. P., 241, 297, 298

Guilford-Schneidman-Zimmerman Interest Survey, 239

- Haganah, Theda, 232, 236
 "Halo effect," 280, 334, 420, 522
 Handwriting, evaluation of, 411-412
 Hardaway, Mathilde, 176
 Hattwick, L. W., 316
 Hawkes, Herbert E., 325
Henmon-Nelson Tests of Mental Ability, 61
 Histogram, 36
 History, testing of, 108, 109, 110
Holtzman Ink Blot Technique, 312
 Home economics, evaluation in, 417
 Hopkins, Kenneth D., 443
- Index of forecasting efficiency, 120
- Individualized instruction
 and educational diagnosis, 458
 effective teaching and, 461
 need for, 461-462
- Industrial arts, ratings in, 416-417
- Instruction
 corrective, 480-482
 materials for, 481
 program for, 482
 emphasis on specific knowledge in, 328-329
 evaluation and, 325-326, 488
 individualized, 458
 need for, 461-462
 improvements in, 321 ff.
 long-range objectives in, 328-329
- Instructional tests, 323-325
 criteria for, 324
- Intelligence quotient
 computation of, 214-216
 constancy of, 216
 origin of, 184
 reasons for variability in, 218-219
- Intelligence tests, 99-100, 119, 130, 132, 138
 development of, 183-193
 effect of cultural background on results from, 217
 group versus individual, 194
 multiscore, 198
 compared with aptitude test batteries, 198-203
 nonverbal, 192-194
 origin of, 183-184
 purpose of, 212-213
 reasons for varying results in, 218-219
 variations in content of, 218
- Interest inventory, 228, 229, 230
 based on empirical study, 234-237
 based on factor analysis, 239
 based on unitary traits, 237-238
 basic interest groups, 241-242
 interpretation of results from, 246-248
 occupational, 240
 predictions from
 on academic achievement, 241
 on occupational satisfaction, 243-244
 for vocational training, 244-246
 types of, 233-240
 validity of, 241-246
- Interests
 aptitudes and, 230-232
 expressed, 229
 factors affecting development of, 228-229
 manifest, 229
 measurement of, 228 ff., 497
 methods of obtaining data on, 229-230
 stability of vocational, 233
 tested, 229
- Internal-consistency method, 86-87, 91, 93, 94
- Interval scales, 64
- Interview, with pupil
 importance of rapport in, 266-267
 interpretation of data from, 267-268
 on personal-social adjustment, 265-268
 preparations for, 266
- Inventory
 of attitudes, 250
 interest. *See* Interest inventory
 personality, 297 ff.
- Iowa Algebra Aptitude Test*, 211

- Iowa Tests of Basic Skills*, 440, 492, 503
- Iowa Tests of Educational Development* (ITED), 444, 445, 513, 550
- Item analysis, 354–357, 478–479
- K-score, 56, 303
- Karnes, M. Ray, 176, 416
- Katz, Martin, 357
- Kaulfers, W. V., 373
- Kawin, Ethel, 508
- Kelley, T. L., 96
- Kent-Rosanoff Free Association Test*, 312
- Klausmeier, Herbert J., 536
- Knowledge
 - of abstractions, 376–377
 - of categories, 375
 - of conventions, 371–373
 - of criteria, 375–376
 - definition of, 367
 - as educational objective, 364
 - performance in skills and, 401
 - of specifics, 368–371
 - of terminology, 368–370
 - of trends and sequences in, 373–374
 - of universals, 376–377
- Kreidt, P. H., 237
- Kruglak, H., 405
- Kuder, G. Frederic, 230, 234
- Kuder Occupational, Form D*, 239
- Kuder Preference Record*, 169, 230, 231, 233, 236, 240, 254, 550, 552
- Kuder Preference Record, Personal*, 239
- Kuder Preference Record, Vocational*, 238, 247, 310
- Kuder-Richardson method, 87, 88, 90, 91, 93, 94, 172, 305
- Kuhlmann-Anderson Intelligence Test*, 503
- Languages, foreign
 - achievement tests for, 447
 - aptitude tests for, 211
- Learning difficulties
 - causes of, 475–477
 - corrective instruction for, 480–482
 - determining, 471–475
 - emotional factors and, 477
- Lee, E. A., 240
- Likert, R., 252
- Likert method of attitude-scale construction, 252–253
- Lindsay, Alexander D., 388
- Linear standard scores, 35, 43
- Linear transformations, 35, 38
- Loevinger, Jane, 308
- Lorge-Thorndike Intelligence Tests*, 99, 130, 131, 192
- McArthur, C., 236
- McCully, C. Harold, 244
- Machover, Karen, 314
- Maier, Thomas, 176
- Manual dexterity, aptitude in, 205, 208
- Marston, William, 297
- Matching questions, 342–345, 368
- Mathematics
 - aptitude tests for, 211
 - grading in, 529–530
 - prognostic tests in, 211
 - testing of, 446
- Maturity, in personality development, 258–259
- Maurer, Katherine M., 115
- Mean, 22
 - computation of, 19, 24, 25, 27, 623
- Measurement
 - of aptitudes. *See* Aptitudes
 - data from. *See* Data
 - definition of, 5
 - direct versus indirect, 150–153
 - educational diagnosis and, 458–462
 - effective teaching and, 461
 - principles of, and test selection, 149 ff.
 - problems in, 7–14
 - relation to evaluation of, 6
 - see also* Evaluation, Tests
- Meier Art Tests*, 204, 209

- Melville, Donald S., 245
- Mental abilities tests. *See* Intelligence tests
- Mental age, 56, 184, 214
computation of, 215
- Mental health. *See* Personal-social adjustment
- Mental Measurement Yearbooks*, 174, 175
- Metropolitan Achievement Test*, 62, 169, 440, 444, 464
- Metropolitan Readiness Tests*, 211
- Micheels, William J., 176, 416
- Michigan Vocabulary Profile Test*, 229
- Minnesota Clerical Test*, 59, 204, 208
- Minnesota Counseling Inventory* (MCI), 309
- Minnesota Multiphasic Personality Inventory* (MMPI), 303, 304, 309
- Minnesota Paper Formboard*, 204, 205, 206
- Minnesota Rate of Manipulation Test*, 208
- Minnesota Spatial Relations Test*, 205, 206
- Modern Democratic State, The*, 388
- Modern Language Aptitude Test*, 211
- Mooney Problem Check Lists*, 310
- Multiple-choice questions, 339–342, 351, 368
- Multiple regression equation, 547
- Multitrait-multimethod matrix, 138, 140
- Murphy, Gardiner, 311–312
- Murray, Henry A., 310
- Music
aptitude in, 209–210
skill in, 421
- Musical Aptitude Test*, 210
- National Guidance Testing Program, 168
- Neurotic, 260–261
- Nominal numbers, 63
- Normal behavior, 259
- Normal distribution curve, 22, 27
characteristics of, 34–36
- Normalized standard scores, 34–40, 43
relationship between types of, 43, 617 ff.
- Norming process, of testing,
definition of, 150
see also Standardized tests
- Norms, 158, 172
age, 20, 55–56
grade placement, 20, 50–55, 165, 439
homogeneous versus population in general, 57–58
importance of, 57, 59
item, 433, 453–454
local, 19
modal-age grade, 62
national, 60, 62, 165, 166
percentile, 165, 166, 439
procedures for obtaining, 60
see also Converted scores
- Number systems, types of, 63
- Objectives
and instruction, 328–329
taxonomy of educational, 363 ff.
- Objective tests, 330, 332
- Objectivity
in evaluation, 460
of teacher-made tests, 331–332
in scoring, 167–168, 422, 429, 430, 522
- Observation
of attitudes, 248–251
of behavior, 270–276
informal, 270–272
recording data from, 272–274
by situational tests, 274–276
systematic, 270
- O'Connor Finger and Tweezer Dexterity Tests*, 208
- Occupational Interest Inventory*, 240, 241

- Oden, M. H., 245
 Ogive, 28
 Ordinal scales, 63
 Orleans Algebra Prognosis Test, 211
 Otis Normal Percentile Chart, 39, 40, 41
Otis Quick-Scoring Mental Ability Tests, 219, 469
- Paranoid, 304
 Parents
 school records and, 508
 teacher conference with, 516-519, 520
Peabody Library Information Test, 492
 Pearson product-moment method, 77-78, 232, 624-626
 Peer-nomination technique, 139, 276, 289-291
 Percentile point, 33
 Percentile rank, 18, 33, 39
 Percentile scores, 28
 advantages of, 30
 computation of, 32
 disadvantages of, 30, 32
 Performance tests
 disadvantages of, 404-405
 scoring of
 by process, 405-406, 408, 411
 by product, 405-406, 409, 411-412
 by ranking, 406-407, 422
 selection of, 403-404
 standardization of conditions for, 404
 types of, 402
 see also Achievement tests, Skills
Personal and Social Development Program, 277
 Personal-social adjustment
 criteria of, 261
 definition of, 259
 evaluation of
 difficulties in, 257
 by informal observation, 270-274
 by opinions of others, 264-265, 276 ff.
 by projective tests, 264, 295-296, 311, 315
 by self-report, 263, 265-269
 by situational tests, 274-276
 by sociometric techniques, 281-289
 by systematic observation, 270
 internal conflicts and, 260
 major concepts of, 258-259
 peer judgment of, 289-291
 teacher rating scales for, 276-280
- Personality
 clinical study of, 262, 311-316
 description of, 261-263
 psychometric study of, 261-262
 see also Personal-social adjustment
 Personality development
 maturity in, 258-259
 normality in, 259
 study of, 257 ff.
 Personality inventory, 297 ff.
 disadvantages of, 299-301
 interpretation of results from, 306-311
 problems checklist as, 310-311
 reliability of, 305-306
 validity of, 301-305
- Personality traits
 homogeneous, 297
 source, 298
 surface, 297
- Personnel and Guidance Journal*, 175
 Personnel Psychology, 175
 Physical Science Study Committee, 447
 Pintner intelligence tests, 192, 469
Pintner-Paterson Performance Scale, 184
 Prediction equations, 547-549
 Prescott, George A., 450
Primary Mental Abilities Test, 91
 Problem-solving questions, 345-347

- Problems checklist, 310–311
- Product scale, 411–412, 413, 422
- Profile, 229, 238, 550
 - analysis of, 556–559
 - aptitude test, 196, 553
 - preparation of, 555
- Project Talent, 61, 62
- Psychologist, evaluation program and, 494
- Psychological tests. *See* Tests, psychological
- Psychometric approach, 261–262, 297 ff.
 - vs. clinical approach, 295–296
- Psychotic, 260
- Purdue Pegboard, 208

- Rank-difference method, 75–76
- Ranking
 - for grading, 406–407, 422
 - teacher-made tests for, 322–323
- Rapport, 266–267
- Rating scale, 406–411, 422
 - advantages of, 514
 - design of, 278–280
 - graphic, 278
 - teacher, 276–280
- Ratio scales, 64
- Raw scores, 17
 - converted to age and grade scores, 43 ff.
 - converted to normalized standard scores, 36, 37
 - converted to percentile scores, 28
- Readiness
 - for arithmetic, 211
 - for reading, 210, 211
- Reading
 - diagnostic tests in, 473–475
 - remedial, 474
- Reading Comprehension Grade Placement (RCGP), 467
- Reading-readiness tests, 119, 153, 210, 211, 495
- Redl, Fritz, 260
- Regression, 82
- Regression effect, 450
- Relevance, of tests, 103–106, 159
- Reliability, 68 ff., 158, 159, 163, 165, 172
 - coefficient of
 - comparison of standard error with, 83
 - definition of, 73
 - factors affecting size of, 93–95
 - methods of obtaining, 84–92
 - standards for, 95–96
 - correlation and, 74
 - definition of, 68
 - of difference scores, 92, 163
 - in evaluating skills, 421–424
 - homogeneity of sample and, 90, 94
 - methods of estimating, 93
 - methods of increasing, 97–99
 - minimum coefficient of, 96
 - of personality inventories, 305–306
 - of teacher-made tests, 322–323
 - validity and, 103
- Remmers, H. H., 333
- Report cards, 512–515
- Review of Educational Research*, 175
- Rice, J. M., 429
- Rimland, Bernard, 300
- Rorschach Ink Blot Test*, 312, 313
- Rosenzweig Picture-Frustration Study*, 313
- Rothney, John W. M., 175, 541
- Rotter Incomplete Sentence Test*, 313

- Sample
 - homogeneity of, 90
 - random, 104
 - validity and, 107
 - representativeness of, 5, 6, 15, 60
 - validity and, 108–109
 - size of, 6, 97, 104, 422
 - in tests, 4
- Sapon, W., 211
- Sarbin, T. R., 243
- Sax, Gilbert, 443, 519
- Scannell, Dale P., 525, 529
- Scatter diagram, 79, 80

- Scholastic Aptitude Test (SAT)*, 170, 171, 172, 217
- School and College Ability Tests (SCAT)*, 168, 171, 438, 439
- Schools, administrative problems of, 9-14
- Science, in achievement tests, 438, 446
- Science Research Associates (SRA) Achievement Series*, 54, 134, 141, 473
- Science Research Associates (SRA) Junior Inventory*, 311
- Science Research Associates (SRA) Youth Inventory*, 310
- Scores
 ability, 553
 composite, 522-525
 difference, 92
 expectancy, 62
 perfect, 18
 true, 71, 72
see also Converted scores, Age scores, Grade scores, Normalized standard scores, Percentile scores, Raw scores, Standard scores, t-scores, Test scores, z-scores
- Scoreze, 168, 439, 471
- Scoring
 ease of, 167-168
 of essay tests, 330, 334
 in evaluative process, 7
 interpretation of, 169
 of objective tests, 330, 332
 objectivity in, 97-98, 167-168, 422, 429, 430, 522
 of performance tests, 405-412
 of standardized tests, 503-504
- Seashore Measures of Musical Talents*, 204, 209
- Segel, David, 232
- Self-report techniques, 263, 265-269
- Sequential Tests of Educational Progress (STEP)*, 99, 162, 163, 168, 169, 413, 438, 439, 440, 444, 445
- Shorthand, aptitude tests for, 211-212
- Simon, Theodore, 184
- Skewed distribution, 27, 37
- Skills
 athletic, 417, 419-420
 basic, 476, 496
 communication, 412-416
 computational, 473
 evaluation of, 401 ff.
 knowledge and, 401
 manipulative, 416-417
 methods for testing, 403-404
 reliability in evaluating, 421-424
 validity in evaluating, 420-421
 work-study, 476
- Snellen Chart, 477
- Social studies
 in achievement test, 438-439, 447
 objective in learning, 328-329
 testing of, 373-374
- Sociogram, 283-286
see also Sociometric techniques
- Sociometric techniques, 281-292
 administration of, 282-283
 interpretation of, 284-286
 peer judgments and, 289-291
 purpose of, 291
 problems in using, 291-292
 reliability of, 286-287
 role of teacher and, 282-283
 selection of questions in, 281-282
 validity of, 286-289
- Spearman rank-difference method, 75-76
- Spearman-Brown formula, 87, 622
- Spelling
 testing of, 98, 107, 152
 use of teaching machines for, 11, 14
- Spencer, Douglas, 297
- Spitzer Study Skills Test*, 492
- Split-halves method, 86, 93, 305
- Stability, measurement of, 90

- Standard deviation, 20, 22
 - basic formula for, 71
 - computation of, 24, 71
 - use of, 21
 - variance and, 71
- Standard error, 73
 - reliability coefficients compared with, 83
- Standard error of estimate
 - correlation and, 82
 - predictive validity coefficient and, 120
- Standard scores, 20 ff., 617–621
 - see also* t-scores, z-scores
- Standardized tests
 - administration of, 500–502
 - characteristics of, 149–150
 - criticisms of, 428, 430
 - evaluation of, 157–166, 170–174
 - preparation for, 500
 - scheduling of, 501–502
 - scoring of, 503–504
 - selection of, 493–494
 - uses of, 432–437, 460
- Stanford Achievement Test*, 51, 54, 56, 61, 62, 429, 443, 444
- Stanford-Binet Scales*, 45, 98, 131, 184, 185, 186–192
 - compared with *Wechsler Intelligence Tests*, 186–190
- Stanine scores, 41, 42, 464, 517, 523
 - advantages of, 42, 43
 - computation of, 41
- Stern, William, 184
- Stevens, Lucia B., 236
- Stone Reasoning Test in Arithmetic*, 429
- Strong, E. K., 233, 234, 243, 244
- Strong Vocational Interest Blank (SVIB)*, 235–237, 239, 241, 244, 245, 247, 304
- Super, Donald E., 175, 200, 208, 241, 254, 539
- Sweden, grading system in, 524
- Survey of Study Habits and Attitudes*, 492
- Synthesis, as educational objective, 366, 388–391
- Szondi Test*, 314
- T-scores
 - comparison of z-scores with, 27
 - computation of, 27
- T-scaled scores, 38, 39, 43
- Taxonomy, of educational objectives, 363 ff.
- Taylor, K., 233
- Teacher-made tests, 321 ff., 492
 - analysis of results from, 354–357
 - content validity of, 325–328
 - content versus goal emphasis in, 328–330
 - directions for, 350–353
 - editing of, 349–350
 - evaluation of, 347–349
 - grouping of items in, 350
 - improvement of, 357–360
 - for instructional purposes, 323–325
 - make-up of, 353
 - preparation for, 349–354
 - for ranking students, 322–323
 - requirements for, 322
- Teachers
 - educational diagnosis and, 459
 - evaluation program and, 321 ff., 487
 - interview with pupil and, 265–268
 - observation of behavior by, 265–267, 270–276
 - parent conference with, 516–519, 520
 - rating scales of, 276–280
 - role in guidance, 535, 536–538
 - role during testing, 502–503
 - study of personal-social adjustment by, 257, 264
 - use of achievement tests by, 432–433
 - use of cumulative records by, 510, 511–512
 - use of projective techniques by, 315
 - use of sociometric techniques by, 282–283
 - use of standardized tests by, 460

- Teaching machines, 11, 14
- "Technical Recommendations for Achievement Tests," 55
- Technical Recommendations for Psychological Tests and Diagnostic Techniques*, 106, 119, 147
- Terman, Lewis M., 184, 245
- Terman-McNemar Group Test of Mental Ability*, 216, 469
- Test items
- completion, 334-336
 - essay, 330, 331, 333, 334, 388
 - grouping of, 350
 - matching, 342-345, 368
 - multiple-choice, 339-342, 351, 368
 - problem solving, 345-347
 - true-false, 337-339, 351
- Test-retest method, 85-86, 305
- Test scores
- effect of anxiety on, 172
 - effect of coaching on, 171
 - effect of cultural background on, 172
 - effect of fatigue on, 171
 - effect of practice on, 131
 - reliability of, 97-100
 - sources of variance in, 69-70
- Testing
- analysis and, 110
 - application and, 110
 - comprehension and, 109
 - evaluation and, 110
 - knowledge and, 109
- Testing program. *See* Evaluation, program of
- Tests
- abilities, 154
 - achievement. *See* Achievement tests
 - administration of, 98, 150, 158, 166-170
 - aptitude. *See* Aptitude tests
 - classification of
 - based on content, 154
 - based on procedure, 153-154
 - by degree of indirectness, 150-153
 - cost of administering, 169-170
 - diagnostic, 471-472
 - essay, 330, 331, 333-334, 388
 - group, 153
 - "identical-element," 152
 - instructional, 323-325
 - intelligence. *See* Intelligence tests
 - length of, 97
 - as measurement of criterion behavior, 103-105
 - moral knowledge, 138
 - objective. *See* Objective tests
 - pencil-and-paper, 153
 - performance. *See* Performance tests
 - power, 153
 - prognostic, 210-212, 234, 496
 - projective, 264, 295-296, 311, 315
 - psychological, 5
 - purposes of, 103, 106-107
 - reading-readiness, 119, 153, 210, 211
 - "related behavior," 152, 402
 - scores of. *See* Test scores
 - selection of, 11-12, 14, 99
 - self-descriptive, 156
 - situational, 274-276
 - space and number, 91
 - speed, 153
 - standardized. *See* Standardized tests
 - teacher-made. *See* Teacher-made tests
 - "verbalized behavior," 153
 - verbal reasoning, 91-92
 - vocabulary, 98
 - "work-sample," 152, 402
- Tests of Primary Mental Abilities*, 198, 199
- Thematic Apperception Test (TAT)*, 313
- Thompson, Anton, 503
- Thompson, George C., 290
- Thorndike, E. L., 138, 411, 429
- Thorndike Handwriting Scale*, 429
- Thorpe, Louis P., 210, 240
- Thurstone, L. L., 194, 198, 251
- Thurstone method of attitude-scale construction, 251-253

- Tomkins-Horn Picture Arrangements Test (PAT)*, 314
- Torgerson, T. L., 419
- Traits
 definition of, 128, 129
 personal, 277
 social, 277
 validity of, 137
 see also Personality traits
- Traxler, A. E., 423
- Treacy, John P., 473
- True-false questions, 337-339, 351
- Tryon, Caroline, 289
- Tyler-Kimber Study Skills Test*, 492
- United States Bureau of the Census, 61
- United States Employment Service, 221, 240
- Validity, 103 ff., 159
 coefficient of, 120-123, 162
 concurrent, 107, 114-119, 130, 146, 157, 160, 162, 171, 302, 304
 construct, 107, 128-143, 161, 162, 302
 content, 107-114, 130, 146, 157, 160, 171, 325
 convergent, 138, 145
 definition of, 103, 145
 discriminant, 138, 145
 in evaluating skills, 420-421
 face, 160
 factorial, 134, 147
 of interest inventories, 241
 of personality inventories, 301-305
 predictive, 107, 119-128, 130, 146, 157, 160, 162, 171, 304
 relevance and, 103-106
 reliability and, 103
- Variance
 computation of, 71
 correlation and, 82
 error, 72, 132
 lasting general, 69-70, 84-85
 lasting specific, 70, 84-85
 sources of, 69-74
 standard deviation and, 71
 temporary general, 70, 84-85
 temporary specific, 70, 84-85
 total, 72
 true, 72, 73
- Verbal-educational factor, in testing, 156, 157
- Vernon, P. E., 241
- Veterans' Administration, vocational guidance of, 545
- Vocabulary, testing of, 98
- Vocation
 measurement of interests in, 236-237
 stability of interests in, 233
 see also Guidance
- Wagner, Eva Bond, 175
- Wattenberg, William W., 260
- Wechsler, David, 185
- Wechsler Adult Intelligence Scale (WAIS)*, 186
 compared with *Stanford-Binet Scales*, 186-190
- Wechsler Intelligence Scale for Children (WISC)*, 186
 compared with *Stanford-Binet Scales*, 186-190
- Wechsler-Bellevue Scale for Adolescents and Adults*, 185, 186
- Wesman, Alexander G., 231
- West, J. Y., 249
- Wetzel "grid," 477
- Whistler, Harry S., 210
- Whitney, A. P., 287
- Wiener, Daniel N., 302
- Wing Standardized Tests of Musical Intelligence*, 210
- Wrenn Study Habits Inventory*, 492
- Wrinkle, William, 491, 514, 520
- z-scores
 comparison of t-scores with, 27
 computation of, 23, 26, 38
 Pearson product-moment method and, 77-78